

# Greedy adaptive walks on a correlated fitness landscape

Su-Chan Park<sup>a,\*</sup>, Johannes Neidhart<sup>b</sup>, Joachim Krug<sup>b</sup>

<sup>a</sup>Department of Physics, The Catholic University of Korea, Bucheon 14662, Republic of Korea

<sup>b</sup>Institut für Theoretische Physik, Universität zu Köln, 50937 Köln, Germany

## Abstract

We study adaptation of a haploid asexual population on a fitness landscape defined over binary genotype sequences of length  $L$ . We consider greedy adaptive walks in which the population moves to the fittest among all single mutant neighbors of the current genotype until a local fitness maximum is reached. The landscape is of the rough mount Fuji type, which means that the fitness value assigned to a sequence is the sum of a random and a deterministic component. The random components are independent and identically distributed random variables, and the deterministic component varies linearly with the distance to a reference sequence. The deterministic fitness gradient  $c$  is a parameter that interpolates between the limits of an uncorrelated random landscape ( $c = 0$ ) and an effectively additive landscape ( $c \rightarrow \infty$ ). When the random fitness component is chosen from the Gumbel distribution, explicit expressions for the distribution of the number of steps taken by the greedy walk are obtained, and it is shown that the walk length varies non-monotonically with the strength of the fitness gradient when the starting point is sufficiently close to the reference sequence. Asymptotic results for general distributions of the random fitness component are obtained using extreme value theory, and it is found that the walk length attains a non-trivial limit for  $L \rightarrow \infty$ , different from its values for  $c = 0$  and  $c = \infty$ , if  $c$  is scaled with  $L$  in an appropriate combination.

**Keywords:** Adaptation, Genotype space, Extreme value theory, Asexual population

## 1. Introduction

Ever since the concept of the *fitness landscape* was introduced by Sewall Wright (1932), it has played a central role in evolutionary biology (de Visser and Krug, 2014). Among the different variants of the concept used in the literature, we here restrict ourselves to fitness landscapes that map the genotype space into the real numbers by assigning a fitness value to every genotype. With this definition, the fitness landscape provides an intuitive picture of evolution as a hill-climbing process. A convenient choice for the genotype space is the  $L$ -dimensional hypercube  $\{0, 1\}^L$ , which contains all binary sequences  $C = (1, 0, 1, \dots, 1, 1)$  of length  $L$ . Rather than specifying the genome on the level of DNA base pairs, the binary sequences keep track of the presence or absence of mutations compared to a wild-type genome, or (in a more coarse-grained representation) the presence or absence of entire genes.

In addition to the underlying fitness landscape, the dynamics of adaptation is governed by the population size  $N$  and the mutation rate  $U$  per genome, both of which are to be compared to the scale of fitness differences summarized in a typical selection coefficient  $s$ . In the *strong selection / weak mutation* (SSWM) regime characterized by the conditions  $Ns \gg 1$  and  $NU \ll 1$  the population is monomorphic for most of the time, and the adaptive process is guided by the landscape structure in a simple way (Gillespie, 1983, 1984). If a mutation to a

fitter genotype occurs it has a nonzero probability of fixing in the population, whereas a mutation to a sequence with lower fitness is certain to go extinct. The low mutation rate makes it very unlikely for double mutations to occur. Accordingly, in this regime the population behaves as a point in sequence space that moves uphill in the fitness landscape by single mutational steps, a process referred to as an *adaptive walk* (Gillespie, 1983, 1984; Kauffman and Levin, 1987). An obvious feature of adaptive walks is that they end on a fitness maximum, that is, a genotype without fitter one-mutant neighbors. This makes the *walk length*, the number of steps until a maximum is reached, a property of interest.

A simplified version of the adaptive walk problem where the effect of mutant fitness on the fixation probability of beneficial mutations is neglected and any neighboring genotype of higher fitness can fix with equal probability was studied by Macken and Perelson (1989) and Flyvbjerg and Lautrup (1992). For rugged landscapes without fitness correlations the mean number of steps of such ‘random’ adaptive walks was found to be of the order of  $\ln L$ . When the effect of the fixation probability is incorporated the mean walk length is still logarithmic in the number of loci, but the coefficient of  $\ln L$  becomes dependent on the distribution of fitness values (Gillespie, 1983; Orr, 2002; Neidhart and Krug, 2011; Jain, 2011; Seetharaman and Jain, 2011). If the infinite  $L$  limit is taken the walks no longer terminate and adaptation can be studied through the unbounded increase of the mean fitness of the population (Park and Krug, 2008).

When the population size is increased beyond the SSWM

\*Tel: +82-2-2164-4524, Fax: +82-2-2164-4764

Email address: spark0@catholic.ac.kr (Su-Chan Park)

regime, the number of segregating sites becomes larger than two. In asexual populations this implies that two beneficial mutations compete with each other for fixation and the one with the larger fitness will be fixed preferentially. This phenomenon is connected to the Hill-Robertson effect (Hill and Robertson, 1966) and is commonly known as *clonal interference* (Gerrish and Lenski, 1998; Wilke, 2004; Park and Krug, 2007; Desai and Fisher, 2007; Park et al., 2010). A rough criterion for the clonal interference regime is provided by the condition  $NU \ln N \gg 1$ . If we denote the mean fixation time in this regime by  $T_0$  (which depends on  $N$ ,  $U$ , and  $s$ ), almost all beneficial single mutant neighbors of the most populated genotype will be present during the fixation process if  $NU T_0 \gg L$ . To model this regime by an adaptive walk, we use a deterministic rule for the next step: the walker chooses the genotype with the largest fitness among the sequences that are one mutation away. This kind of adaptive walk was called a ‘perfect’ or ‘gradient’ adaptive walk by Orr (2002, 2003), but here we follow Kauffman and Levin (1987) in referring to it as a *greedy* walk. Orr (2003) calculated the length of a greedy adaptive walk on an uncorrelated fitness landscape using an order statistics approach that is independent of the fitness distribution, provided it is continuous. In the limit  $L \rightarrow \infty$  the mean walk length is given by  $e - 1 \approx 1.72$ , which was suggested to be a lower bound on the mean number of steps for any adaptive walk [see also Rosenberg (2005)]. Note that for this description to faithfully represent adaptation under strong clonal interference, the mutation rate has to be small enough such that the creation of double mutants can be neglected (Szendro et al., 2013a).

The studies of adaptive walks mentioned above were based on the assumption of an uncorrelated random fitness landscape with maximal ruggedness, which is not supported by empirical evidence (Miller et al., 2011; Szendro et al., 2013b; de Visser and Krug, 2014). The effect of fitness correlations on adaptive walks has so far been addressed mostly in the context of ‘block model’ landscapes in which the genotype is subdivided into independent modules, each of which is assigned a random fitness, and the mean walk length is additive over modules (Perelson and Macken, 1995; Orr, 2006; Seetharaman and Jain, 2014; Nowak and Krug, 2015). Here we consider greedy adaptive walks on another class of tunably rugged fitness landscapes, the rough mount Fuji (RMF) model, which was originally introduced in the context of protein evolution (Aita et al., 2000). In the RMF model an uncorrelated random fitness landscape is superimposed on a linear fitness gradient, and the slope of this gradient serves as a tuning parameter controlling the ruggedness of the landscape.

The RMF model has recently been found to provide a convenient parametrization of many empirical fitness data sets (Franke et al., 2011; Szendro et al., 2013b; Neidhart et al., 2013), while at the same time allowing for detailed mathematical analysis of a wide range of landscape properties (Neidhart et al., 2014; Park et al., 2015). Of particular interest for our work are the results on the existence of selectively accessible mutational pathways, defined here as pathways to the global fitness maximum along which fit-

ness increases monotonically and which are moreover *directed*, in the sense that the distance to the global optimum decreases in each step (Weinreich et al., 2005; Franke et al., 2011). Hegarty and Martinsson (2014) have shown that such pathways exist in the RMF model with a probability approaching unity for  $L \rightarrow \infty$ , whereas this probability tends to zero for uncorrelated landscapes. A population following a directed accessible pathway would perform an adaptive walk of  $O(L)$  steps, much longer than the walks on uncorrelated landscapes. However, the biological significance of accessible paths is not evident, because an evolving population may not find them even if they exist (Szendro et al., 2013a; Park et al., 2015).

In this paper, we study greedy adaptive walks on the RMF fitness landscape, focusing on the mean number of steps when  $L$  is very large. For a specific choice of the distribution of the random fitness component in the RMF model we obtain an analytic solution for the full distribution of walk lengths and show that it attains a non-degenerate limit for  $L \rightarrow \infty$ , similar to Orr’s analysis of the uncorrelated case (Orr, 2003). We also consider the dependence of the walk length on the distance of the initial genotype from the *reference state*, and show that in a range of distances the walk length varies non-monotonically with the strength of the fitness gradient.

Arbitrary distributions of the random fitness component can be treated in the limit  $L \rightarrow \infty$  by exploiting the convergence of the maximum of  $L$  random variables to one of the universal distributions of extreme value theory (EVT) (de Haan and Ferreira, 2006). The EVT approach to adaptation was pioneered by Gillespie (1984) and Orr (2002) and has meanwhile become an established conceptual framework that allows to organize and quantify the relation between the distribution of mutational effects and the corresponding adaptive behavior (Joyce et al., 2008; Orr, 2010; Rokytka et al., 2008; Schenk et al., 2012; Bank et al., 2014). Similar to the analysis of fitness landscape properties for the RMF model presented by Neidhart et al. (2014), we find that the behavior of the walk length is governed by the interplay between the ruggedness parameter and the tail properties of the distribution of the random fitness component. Specifically, if the tail of the distribution is fatter than exponential, the walk length reverts to the behavior found by Orr for uncorrelated landscapes for any fixed value of the fitness gradient. On the other hand, for tails thinner than exponential the effective strength of the fitness gradient increases without bound with increasing  $L$ , such that the greedy walks traverse the entire landscape with high probability for  $L \rightarrow \infty$ . A non-trivial limit of the walk length is attained only when  $c$  and  $L$  are scaled together in a particular combination.

## 2. Definitions

The RMF fitness landscape is constructed from an additive ‘mount Fuji’ fitness landscape by adding an independent and identically distributed (i.i.d.) random variable to the fitness of every genotype. By  $C$  we denote a binary sequence of length  $L$  which represents the genotype. In particular, we will call the sequence  $C_r = (1, 1, \dots, 1)$  the *reference sequence* which has the largest fitness in the purely additive landscape. Its antipodal

point on the hypercube, the sequence with all elements 0, will be denoted by  $C_a$ . The fitness of a sequence  $C$  in the RMF fitness landscape is then assigned as

$$W(C) = -cd_r(C) + \xi_C, \quad (1)$$

where  $d_r(C)$  is the Hamming distance between  $C$  and the reference sequence  $C_r$ ,  $c$  is a positive real number, and  $\{\xi_C\}$  are i.i.d. random variables with probability density  $f(\xi)$  and cumulative distribution function  $F(\xi)$ , defined as

$$F(\xi) = \int_{-\infty}^{\xi} f(x) dx. \quad (2)$$

The definition (1) should be interpreted in the Malthusian sense, where fitness values can be positive or negative. What Hegarty and Martinsson (2014) proved is that for  $c > 0$  in the limit  $L \rightarrow \infty$  there is almost surely a directed path from the antipode  $C_a$  to the reference sequence  $C_r$  along which fitness is monotonically increasing, irrespective of the actual form of  $f(\xi)$ , whereas for  $c = 0$  such paths almost surely do not exist.

Since we are interested in greedy walks, the statistics of the maximal value among groups of i.i.d. random variables will play an important role. For this reason we introduce the probability  $G_k(x)$  that the largest value among  $L - k + 1$  ( $k \geq 1$ ) i.i.d.  $\xi$ 's is smaller than  $x$ , which is

$$G_k(x) \equiv \left( \int_{-\infty}^x f(y) dy \right)^{L-k+1} = F(x)^{L-k+1}, \quad (3)$$

with the corresponding density  $g_k(x)$

$$g_k(x) = (L - k + 1)f(x)F(x)^{L-k}. \quad (4)$$

The reason for considering  $L - k + 1$  variables rather than  $k$  variables will become clear in Sec. 3.

As has been noted previously (Franke et al., 2010, 2011; Neidhart et al., 2014), many properties of the RMF model take on a particularly simple form when the random variables  $\xi_C$  are drawn from the Gumbel distribution  $f(x) = e^{-x-e^{-x}}$ , and we will adopt this choice in Sec. 3. For the Gumbel distribution,  $G_k(x)$  and  $g_k(x)$  become

$$g_k(x) = (L - k + 1)e^{-x-(L-k+1)e^{-x}}, \quad (5)$$

$$G_k(x) = \int_{-\infty}^x g_k(x) = e^{-(L-k+1)e^{-x}}. \quad (6)$$

The Gumbel distribution is one of the three universal limiting distributions that arise in extreme value theory (de Haan and Ferreira, 2006), and we will exploit this connection in Sec. 4 where we study the properties of greedy adaptive walks for general choices of the distribution  $f(x)$ .

### 3. Gumbel-distributed random fitness component

#### 3.1. Greedy walks starting from the antipodal sequence

Our analysis begins with the greedy walk starting from the antipodal sequence  $C_a$ . As mentioned before, the probability

that at least one accessible path from  $C_a$  to  $C_r$  exists converges to unity as  $L \rightarrow \infty$  for any finite  $c > 0$  (Hegarty and Martinsson, 2014). If the greedy walker takes such a path with probability of  $O(1)$ , the mean number of steps will be  $O(L)$ . On the other hand, the RMF with  $c = 0$  is identical to the uncorrelated rugged landscape or the House-of-Cards model (Kingman, 1978) and the mean number of steps of greedy walks is  $e - 1 \approx 1.72$  in the limit of infinite  $L$  (Orr, 2003). Thus the first question to address is whether the greedy walk length remains finite for  $L \rightarrow \infty$  when  $c > 0$ .

##### 3.1.1. Exact solution

To find the mean walk distance, we consider the probability  $H_l$  that the walker takes at least  $l$  steps. For convenience, we denote the sequence at the  $l$ -th step by  $C_l$  with  $C_0 = C_a$ . The fitness of  $C_l$  is the largest among the single mutant neighbors of  $C_{l-1}$ . To find  $H_l$ , we make the assumption that  $d_r(C_l)$  is a decreasing function in  $l$ , that is, the walker only takes steps in the direction towards the reference sequence  $C_r$ , referred to as the *uphill* direction in the following. This assumption is plausible if  $L \gg l$ , because a *downhill* step is possible only if the largest among the  $l$  random fitness components of the downhill neighbors exceeds the largest among the  $L - l$  random fitness components of the uphill neighbors by at least  $2c$ . Obviously, for reasonably large  $L$  and a setting with rather short walks, this probability is negligible. The validity of this assumption will be ascertained later in a self-consistent way. Once the  $H_l$  have been determined, it follows that the greedy walk takes exactly  $l$  steps with probability  $H_l - H_{l+1}$  and, in turn, the mean number of steps is

$$\langle l \rangle = \sum_{l=1}^L l(H_l - H_{l+1}) = \sum_{l=1}^L H_l, \quad (7)$$

where  $H_{L+1}$  is set to 0.

Let  $J_l(x)$  be the probability that the walker takes at least  $l$  steps with  $W(C_l) < -c(L - l) + x$  and let  $j_l(x) = \frac{d}{dx} J_l(x)$  ( $l = 0, 1, \dots, L$ ). Obviously,

$$H_l = \lim_{x \rightarrow \infty} J_l(x) = \int_{-\infty}^{\infty} j_l(y) dy. \quad (8)$$

A recursion relation for  $j_l(x)$  can be derived immediately from the definition:

$$j_l(x) = g_l(x)J_{l-1}(x + c) = g_l(x) \int_{-\infty}^{x+c} dy j_{l-1}(y) \quad (9)$$

with  $j_0(x) = f(x) = e^{-x-e^{-x}}$ . Since  $C_{l-1}$  has  $L - l + 1$  nearest neighbors in the uphill direction, we have considered  $g_l(x)$  defined in Eq. (4) in the recursion relation.

Introducing

$$\begin{aligned} a_k &= e^{-kc} + \sum_{m=0}^{k-1} (L - k + m + 1)e^{-mc} \\ &= \frac{L(1 - e^{-kc}) - e^{-(k+1)c} - k}{1 - e^{-c}} + \frac{1 - e^{-(k+1)c}}{(1 - e^{-c})^2} \end{aligned} \quad (10)$$

which satisfies the recursion relation  $a_{k+1} = (L-k) + e^{-c}a_k$  with  $a_0 = 1$ , we can write

$$j_l(x) = \frac{L!}{(L-l)!} \left( \prod_{k=0}^{l-1} \frac{1}{a_k} \right) e^{-x-a_l e^{-x}}, \quad (11)$$

for  $l \geq 1$ , which can be proved straightforwardly by mathematical induction. Thus, we get

$$H_l = \prod_{k=1}^l \frac{L-k+1}{a_k} \quad (12)$$

as an exact expression for the distribution of walk length. Note that in the above derivation, the sign of  $c$  does not play any role, which implies that the case of negative  $c$  can be studied within the same scheme and Eq. (12) is valid for any  $c$ . By symmetry, a greedy walk with negative  $c$  can be interpreted as a walk starting from the reference sequence  $C_r$  (see Sec. 3.2 for further discussion).

Since it does not appear feasible to extract simple analytic formulae from (12) for arbitrary  $c$  and  $L$ , below we will present approximate calculations for certain limiting cases. Before delving into detail, we derive a simple upper bound on  $\langle l \rangle$ . Since  $a_k \geq (L-k+1) + (L-k+2)e^{-c} \geq (L-k+1)(1+e^{-c})$  for  $k \geq 2$  and  $a_1 = L + e^{-c} \geq L$ ,

$$H_l \leq (1 + e^{-c})^{-(l-1)}, \quad (13)$$

which gives

$$\langle l \rangle = \sum_{l=1}^L H_l \leq \sum_{l=1}^{\infty} (1 + e^{-c})^{-(l-1)} = 1 + e^c. \quad (14)$$

This upper bound clearly shows that  $\langle l \rangle / L \rightarrow 0$  as  $L \rightarrow \infty$  for any  $c$  when  $\xi_C$  is drawn from the Gumbel distribution. That is, it is highly unlikely that a greedy walk can follow an accessible path all the way to the reference state, although such paths exist with probability 1 as shown by Hegarty and Martinsson (2014).

### 3.1.2. The limit $L \rightarrow \infty$ at finite $c$

Since Eq. (13) is valid for any  $L$ ,  $H_l$  should be exponentially small for  $l \sim O(L)$  once  $L \gg e^c$ . This self-consistently affirms the validity of the assumption used in writing down  $H_l$ . In order to extract the  $L \rightarrow \infty$  limit of  $H_l$  from (12), we use  $a_k \sim L(1 - e^{-kc})/(1 - e^{-c})$  to obtain

$$H_l = \prod_{k=1}^l \frac{1 - e^{-c}}{1 - e^{-kc}}. \quad (15)$$

This expression has an appealing interpretation in terms of so-called  $q$ -analogues (Koekoek et al., 2010). Recall that the  $q$ -analogue of a number  $n$  can be defined by  $[n]_q = (1 - q^n)/(1 - q)$ , which satisfies the basic property that  $\lim_{q \rightarrow 1} [n]_q = n$ . Defining the  $q$ -factorial as  $[n]_q! = \prod_{k=1}^n [k]_q$ , we see that  $H_l = ([l]_{e^{-c}}!)^{-1}$  which reduces to Orr's result  $H_l = (l!)^{-1}$  in the limit  $c \rightarrow 0$ ,  $e^{-c} \rightarrow 1$ . Moreover, the mean walk length is given by

$$\langle l \rangle = \sum_{l=1}^{\infty} H_l = \exp_{e^{-c}}(1) - 1, \quad (16)$$

where  $\exp_q(x)$  is the  $q$ -exponential function, defined as

$$\exp_q(x) = \sum_{n=0}^{\infty} \frac{1}{[n]_q!} x^n.$$

In fact an alternative derivation of (15) can be set up in complete analogy to the original approach of Orr (2003) [see Neidhart (2014)].

We note for later reference that the expression (15) has been derived previously for the probability that  $l$  random variables  $y_k = x_k + ck$  are ascendingly ordered,  $y_1 < y_2 < \dots < y_l$ , where the  $x_k$  are drawn independently from a Gumbel distribution (Franke et al., 2010). The reason for this coincidence will become clear below in Sec. 4.1.

### 3.1.3. Approximations for large and small $c$

We next evaluate (15) for large and small  $c$ , respectively. If  $c \gg 1$ ,  $H_l$  can be approximated as

$$H_l \approx \frac{(1 - e^{-c})^{l-1}}{1 - e^{-2c}} \quad (17)$$

for  $l \geq 2$  and  $H_1 = 1$ . In the above approximation, we have kept terms up to  $O(e^{-2c})$  in the denominator. Hence the mean distance becomes

$$\langle l \rangle \approx 1 + \sum_{k=2}^{\infty} \frac{(1 - e^{-c})^{k-1}}{1 - e^{-2c}} = \frac{e^c - e^{-2c}}{1 - e^{-2c}} \approx e^c + e^{-c} \quad (18)$$

which is close to the upper bound of Eq. (14).

For  $|c| \ll 1$ , we expand  $(1 - e^{-c})/(1 - e^{-kc})$  up to  $O(c^3)$ , which yields

$$\frac{1 - e^{-c}}{1 - e^{-kc}} \approx \frac{1}{k} \exp\left(\frac{k-1}{2}c - \frac{k^2-1}{24}c^2 + O(c^4)\right). \quad (19)$$

Accordingly,  $H_l$  is approximated as

$$\begin{aligned} l!H_l &\approx \exp\left(\frac{l(l-1)}{4}c - (2l^3 + 3l^2 - 5l)\frac{c^2}{144} + O(c^4)\right) \\ &= 1 + \frac{(l)_2}{4}c + \frac{9(l)_4 + 32(l)_3}{288}c^2 \\ &\quad + \frac{3(l)_6 + 32(l)_5 + 72(l)_4 - 24(l)_2}{1152}c^3 + O(c^4), \end{aligned} \quad (20)$$

where the Pochhammer symbol  $(l)_k = l!/(l-k)!$  has been used. Since  $\sum_{l=1}^{\infty} (l)_k/l! = e - \delta_{k,0}$ , the mean distance becomes

$$\langle l \rangle = e \left( 1 + \frac{1}{4}c + \frac{41}{288}c^2 + \frac{83}{1152}c^3 \right) - 1, \quad (21)$$

which reproduces the result by Orr (2003) when  $c = 0$ . The fact that the leading order correction is linear in  $c$  implies that walks starting at the reference sequence ( $c < 0$ ) are *shorter* than  $e - 1$  when  $|c|$  is small. We will see below in Sec. 3.2 how this result generalizes to walks starting close to, but not at the reference sequence.



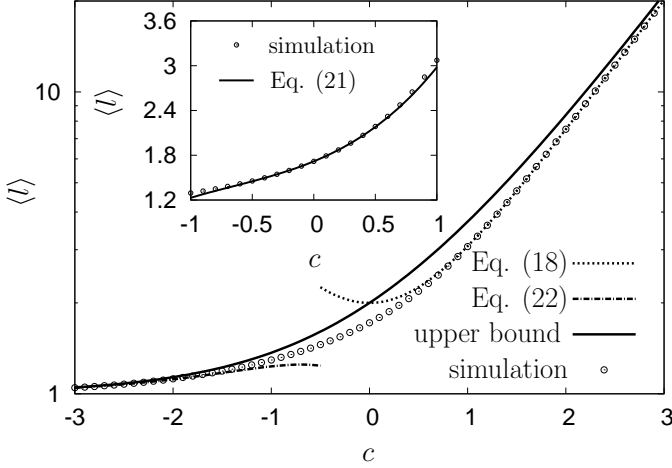


Figure 1: Semi-logarithmic plot of the mean walk length  $\langle l \rangle$  as a function of the strength  $c$  of the fitness gradient. Simulation data are shown together with the approximations Eqs. (18), (22), and the upper bound Eq. (14). Inset: Linear plot of  $\langle l \rangle$  vs.  $c$  together with Eq. (21)

If  $c < 0$  and  $|c| \gg 1$ ,  $(L - k + 1)/a_k \approx e^{-(k-1)|c|}$  and  $H_l = O(e^{-l(l-1)|c|/2})$ . Hence, to keep terms up to order  $O(e^{-2|c|})$ , it is enough to consider only  $H_1 + H_2$ , which gives

$$\langle l \rangle = 1 + e^{-|c|} - e^{-2|c|} + O(e^{-3|c|}). \quad (22)$$

Note that even if  $|c| \rightarrow \infty$ , the walker takes at least one step. This is because we take  $L \rightarrow \infty$  limit before  $|c| \rightarrow \infty$  limit and under this order of limits the probability that the reference sequence is a local maximum is zero for any  $c$ . For later purposes we recall that the probability for a sequence at distance  $d$  from the reference sequence to be a local fitness maximum is given by (Neidhart et al., 2014)

$$p_c^{\max}(d) = \frac{1}{1 + de^c + (L - d)e^{-c}} \quad (23)$$

which vanishes when the limit  $L \rightarrow \infty$  is taken for  $d = 0$  and fixed  $c$ . Thus the walker needs to take at least one step to reach a maximum.

In Fig. 1, we compare  $\langle l \rangle$  obtained from simulations of  $10^8$  independent realization with sequence length  $L = 2^{30}$  to the approximations Eqs. (18), (21), and (22) together with the upper bound of Eq. (14). The simulation method is explained in Appendix A. As a rule of thumb, the large  $|c|$  approximations work well for  $|c| > 1$  and the approximation for  $|c| \ll 1$  becomes accurate for  $|c| < 1$ .

### 3.1.4. The limit $c \rightarrow \infty$ at finite $L$

For finite  $L$ , it is clear that the mean walk length should approach  $L$  as  $c \rightarrow \infty$ . This limit can be attained when  $c$  is much larger than the (typical) largest value among  $L$  i.i.d. random variables. For the Gumbel case, this corresponds to  $\ln L \ll c$  or  $Le^{-c} \ll 1$ . To find an approximate solution of  $\langle l \rangle$  under this condition, we go back to Eq. (12) and expand  $(L - k + 1)/a_k$  in terms of  $e^{-c}$  as

$$\frac{L - k + 1}{a_k} \approx 1 - e^{-c} \left( 1 + \frac{1}{L - k + 1} \right) \quad (24)$$

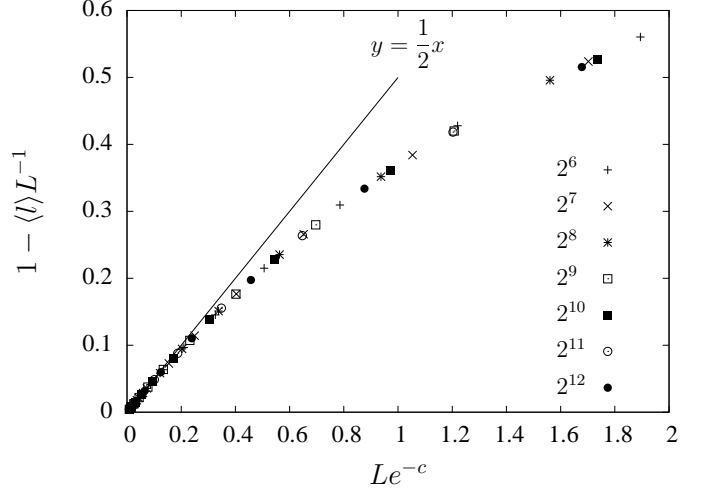


Figure 2: Mean walk length  $\langle l \rangle$  is analyzed for large  $c$  by plotting  $1 - \langle l \rangle L^{-1}$  vs.  $Le^{-c}$  for  $L = 2^6, 2^7, \dots, 2^{12}$ . All data points nicely collapse onto a single curve which is the scaling function  $\Lambda(x)$  in Eq. (27). The asymptotic behavior  $\Lambda(x) \approx x/2$  for  $x \rightarrow 0$  is also confirmed.

for  $k \geq 2$  and  $L/a_1 \approx 1 - e^{-c}/L$ , where we have kept terms up to  $O(e^{-c})$ . Hence

$$\begin{aligned} H_l &\approx 1 - \frac{e^{-c}}{L} - e^{-c} \sum_{k=2}^l \left( 1 + \frac{1}{L - k + 1} \right) \\ &= 1 - e^{-c}(l - 1) - e^{-c} \sum_{k=1}^l \frac{1}{L - k + 1}, \end{aligned} \quad (25)$$

which gives

$$\begin{aligned} \langle l \rangle &\approx L - e^{-c} \left( \frac{L(L - 1)}{2} + \sum_{l=1}^L \sum_{k=1}^l \frac{1}{L - k + 1} \right) \\ &= L \left( 1 - \frac{L + 1}{2} e^{-c} \right). \end{aligned} \quad (26)$$

As anticipated,  $Le^{-c}$  appears as an expansion parameter and  $\langle l \rangle$  approaches  $L$  as  $c \rightarrow \infty$ . Thus, it is quite plausible to assume a scaling form such that

$$1 - \frac{\langle l \rangle}{L} = \Lambda(Le^{-c}), \quad (27)$$

where  $\Lambda(x)$  is a scaling function with asymptotic behavior  $\Lambda(x) \approx x/2$  for sufficiently small  $x$ . That is, if we plot  $1 - \langle l \rangle/L$  as a function of  $Le^{-c}$  for sufficiently large  $L$ , the data obtained for different combinations of  $L$  and  $c$  should collapse onto a single curve. To confirm this, we performed simulations for  $L$  ranging from  $2^6$  to  $2^{12}$ . Figure 2 which is the result of  $10^8$  independent realizations for each data point indeed confirms the existence of such a scaling function.

### 3.2. Greedy walks with arbitrary starting point

In this section, we relax the assumption that the walk always starts at the antipodal sequence  $C_a$  and calculate the mean number of steps in the case that the initial genotype has Hamming

distance  $d_0$  from the reference sequence  $C_r$ . Note that the case treated in the previous section correspond to  $d_0 = L$  and the case with  $c < 0$  in the previous section can be understood as a greedy walk starting at  $d_0 = 0$  with positive  $c$ . We consider the limit  $L, d_0 \rightarrow \infty$  with  $\alpha = d_0/L$  kept finite. Since the RMF landscape is symmetric under the simultaneous transformations  $c \mapsto -c$  and  $d_0 \mapsto L - d_0$ , we can set  $c$  to be non-negative without loss of generality.

### 3.2.1. Exact asymptotic solution

Unlike the previous section, the initial genotype has  $O(L)$  neighbors in both the uphill and downhill directions, and we cannot exclude the possibility that the walker takes a downhill step. Assume that the walker arrives at the sequence  $C_l$  at the  $l$ -th step and  $d_r(C_l) = d$ . Note that  $d_0 - d$  needs not be the same as  $l$ . Now, we introduce the function

$$q(y, \sigma) = \begin{cases} \frac{d(F(y)^d)}{dy} (F(y + 2c))^{L-d} & \text{if } \sigma = +1, \\ \frac{d(F(y)^{L-d})}{dy} (F(y - 2c))^d & \text{if } \sigma = -1, \end{cases} \quad (28)$$

which is interpreted as the probability density that the largest fitness among the uphill (downhill) neighbors has the random contribution  $y$  and all downhill (uphill) neighbors have smaller fitness when  $\sigma = 1$  ( $-1$ ).

As in Sec. 3.1, the probability of taking at least  $l$  steps is denoted by  $H_l$ . Since the walker may move in the uphill or downhill direction with non-negligible probability, we have to take into account all possible combinations of directions. If  $\sigma_l \in \{\pm 1\}$  is the change in the distance from the antipodal sequence  $C_a$  at the  $l$ -th step, then the change in  $d$  over a path is stored in an ordered set  $\{\sigma\}_l = (\sigma_1, \sigma_2, \dots, \sigma_l)$ . Defining  $M_l \equiv \sum_{i=1}^l \sigma_i$ , the Hamming distance from  $C_r$  after  $l$  steps is  $d = d_0 - M_l$ . We assume (and will subsequently verify) that the probability that the walker takes  $O(L)$  steps is exponentially small for large  $L$ . Accordingly, the scaled distance  $d/L$  and therefore the function  $q(y, \sigma)$  in (29) do not change significantly during the walk. Within this assumption, we can approximate (28) in the form

$$q(y, \sigma) = L\beta s_\sigma Q(x + \sigma c), \quad (29)$$

where  $\beta = \alpha e^c + (1 - \alpha)e^{-c}$ ,  $s_1 = \alpha e^c / \beta$ ,  $s_{-1} = 1 - s_1$ , and  $Q(x) = \exp(-x - e^{-x}L\beta)$ , which is independent of  $l$ .

Let  $J_l(x, \{\sigma\}_l)$  be the probability that a walk has moved according to  $\{\sigma\}_l$  and the fitness of the sequence at the  $l$ th step is smaller than  $-c(d_0 - M_l) + x$ . With  $j_l(x, \{\sigma\}_l) = \frac{d}{dx} J_l(x, \{\sigma\}_l)$  we then have, in analogy to (8),

$$H_l = \sum_{\{\sigma\}_l} J_l(\infty, \{\sigma\}_l) = \sum_{\{\sigma\}_l} \int_{-\infty}^{\infty} j_l(x, \{\sigma\}_l) dx, \quad (30)$$

where the summation is over all possible  $2^l$  combinations of  $\{\sigma\}_l$ . Similar to (9) one can construct a recursion relation for  $j_l$ , which reads

$$\begin{aligned} j_l(x, \{\sigma\}_l) &= q(x, \sigma_l) J_{l-1}(x + \sigma_l c, \{\sigma\}_{l-1}) \\ &= q(x, \sigma_l) \int_{-\infty}^{x + \sigma_l c} j_{l-1}(y, \{\sigma\}_{l-1}) dy. \end{aligned} \quad (31)$$

For  $l = 1$ ,

$$\begin{aligned} j_1(x, \{\sigma\}_1) &= q(x, \sigma_1) \int_{-\infty}^{x + \sigma_1 c} f(y) dy \\ &= q(x, \sigma_1) F(x + \sigma_1 c) \approx q(x, \sigma_1), \end{aligned} \quad (32)$$

where we have approximated  $F \approx 1$  because the relevant fitness values reside far in the tail of the distribution when  $L$  is large.

If we neglect the effect of the change in  $d$  on  $q(x, \sigma_l)$  as assumed above, we get

$$\begin{aligned} J_l(\infty, \{\sigma\}_l) &\approx (L\beta)^l \left( \prod_{k=1}^l s_{\sigma_k} \right) \int_{-\infty}^{\infty} dy_l Q(y_l + \sigma_l c) \times \\ &\times \prod_{k=1}^{l-1} \int_{-\infty}^{y_k + \sigma_k c} Q(y_{k-1} + \sigma_{k-1} c) dy_{k-1} \\ &= \prod_{k=1}^l \frac{s_{\sigma_k}}{1 + \sum_{m=1}^{k-1} \exp(-cM_m)}, \end{aligned} \quad (33)$$

where  $\prod'$  in the second line signifies an index-ordered product in descending order of  $k$ , which should be interpreted as 1 if  $l = 1$ . The solvability of the nested chain of integrals in (33) is specific to the Gumbel distribution; see Appendix B. From Eqs. (30) and (33), we arrive at our central result

$$H_l = \sum_{\{\sigma\}_l} \prod_{k=1}^l \frac{s_{\sigma_k}}{1 + \sum_{m=1}^{k-1} \exp(-cM_m)} \quad (34)$$

which reduces to Eq. (15) when  $\alpha = 1$ .

### 3.2.2. Dependence of the walk length on $\alpha$ and $c$

Since  $\exp(-mc) \leq \exp(-cM_m) \leq \exp(mc)$  and  $s_1 + s_{-1} = 1$ , the expression (34) is bounded from below and above by its values for  $\alpha = 0$  and  $\alpha = 1$ , respectively

$$H_l|_{\alpha=0} = \prod_{k=1}^l \frac{1 - e^c}{1 - e^{kc}} \leq H_l \leq \prod_{k=1}^l \frac{1 - e^{-c}}{1 - e^{-kc}} = H_l|_{\alpha=1}. \quad (35)$$

In fact, using  $\frac{d}{d\alpha} s_1 = -\frac{d}{d\alpha} s_{-1} \geq 0$  and  $\exp(-c + A) \leq \exp(c + A)$  for any real  $A$ , one can easily see that  $\frac{d}{d\alpha} H_l \geq 0$ , where the equality holds only when  $c = 0$ . That is,  $H_l$  is an increasing function of  $\alpha$ , and correspondingly the mean walk length (7) decreases monotonically as the position of the starting point approaches the reference sequence, which is easily conceivable.

By contrast, the dependence of the mean walk length on  $c$  is more complex. We have seen above in Sec. 3.1.3 that the walk length decreases with increasing  $c$  when the walk starts at the reference sequence ( $\alpha = 0$ ), and we will now show that such an initial decrease occurs whenever  $\alpha < \frac{1}{2}$ . On the other hand, for very large  $c$  the walk length must approach  $\alpha L$  for any  $\alpha > 0$ , and we must therefore expect a non-monotonic dependence on  $c$  for  $0 < \alpha < \frac{1}{2}$ . Such a behavior was already reported by Neidhart et al. (2014) on the basis of numerical simulations.

When  $c \ll 1$ , we can approximate  $H_l$  up to  $O(c^2)$  as (see Appendix B for the derivation)

$$l! H_l = 1 + \delta \frac{(l)_2}{4} c + \delta^2 c^2 \frac{9(l)_4 + 32(l)_3}{288} + (1 - \delta^2) c^2 \frac{7(l)_3}{108}, \quad (36)$$

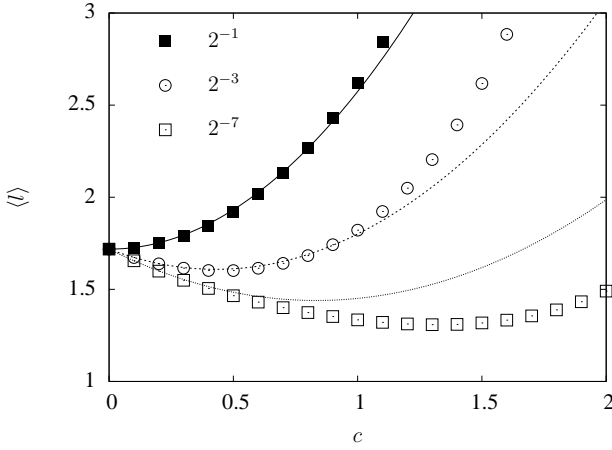


Figure 3: The mean walk length  $\langle l \rangle$  is plotted as a function of  $c$  for different starting points  $\alpha = 2^{-1}$ ,  $2^{-3}$ , and  $2^{-7}$  (from top to bottom) with comparison to the expansion in Eq. (37). The sequence length is  $L = 2^{60}$  and the number of independent runs for each data point is  $10^9$ .

where  $\delta = s_1 - s_{-1} = (\alpha e^c - (1 - \alpha)e^{-c})/\beta$ . Accordingly, the mean number of steps becomes

$$\begin{aligned} \langle l \rangle &\approx e - 1 + \frac{\delta}{4}ec + \frac{41\delta^2}{288}ec^2 + \frac{7}{108}(1 - \delta^2)ec^2 \\ &\approx e - 1 + \frac{2\alpha - 1}{4}ec + \frac{123 + 596\alpha(1 - \alpha)}{864}ec^2, \end{aligned} \quad (37)$$

where we have also expanded  $\delta$  up to  $O(c^2)$ . Hence  $\langle l \rangle$  is an increasing function of  $c$  for  $\alpha \geq \frac{1}{2}$  when  $c$  is small enough, while for  $\alpha < \frac{1}{2}$  the mean walk length initially decreases with  $c$  for small  $c$ . Since the walk length is known to increase at large  $c$ , it follows that there must be a turning point which, in the quadratic approximation (37), is given by

$$c_{\text{turn}} \approx \frac{108(1 - 2\alpha)}{123 + 596\alpha(1 - \alpha)}. \quad (38)$$

A comparison of Eq. (37) with simulations is shown in Fig. 3, which illustrates the accuracy of the analytic expression (37) for small  $c$ . As predicted, it also confirms the absence of a turning point for  $\alpha \geq \frac{1}{2}$ . As  $\alpha$  decreases, the position of the turning point found in the simulations moves to larger  $c$ , which makes the small  $c$  approximation inaccurate for precisely pinpointing  $c_{\text{turn}}$ .

From Fig. 3, the position  $c_{\text{turn}}$  of the turning point seems to diverge as  $\alpha \rightarrow 0$ . When  $\alpha = 0$ , the mean walk length decreases as  $\langle l \rangle \approx 1 + e^{-c}$  for sufficiently large  $c$  as shown in Sec. 3.1.3. When  $\alpha$  is very small,  $\langle l \rangle$  should therefore first decrease as  $1 + e^{-c}$ , but eventually increase with  $c$  for sufficiently large  $c$ . As in the case of  $c < 0$  and  $|c| \gg 1$  in Sec. 3.1.3, when  $\langle l \rangle - 1 \ll 1$  this quantity is expected to be well approximated by

$$\langle l \rangle - 1 \approx H_2 = \frac{s_1}{1 + e^{-c}} + \frac{s_{-1}}{1 + e^c} = \frac{1 + \varepsilon e^{3c}}{(1 + e^c)(1 + \varepsilon e^{2c})}, \quad (39)$$

where  $\varepsilon = \alpha/(1 - \alpha)$ . In Fig. 4, we compare simulation results for small  $\alpha$  ( $\alpha \leq 2^{-10}$ ) with Eq. (39), which shows an excellent

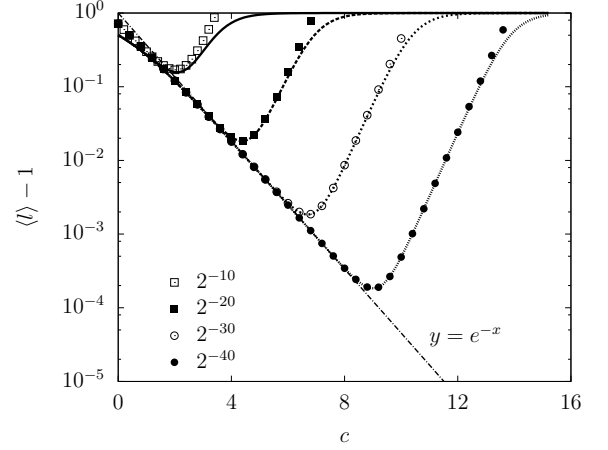


Figure 4: Semi-logarithmic plots of  $\langle l \rangle - 1$  vs.  $c$  for  $\alpha = 2^{-10}$ ,  $2^{-20}$ ,  $2^{-30}$ , and  $2^{-40}$  (from top to bottom) with comparison to Eq. (39). The sequence length is  $L = 2^{60}$  and the number of independent runs for each data point is  $10^9$ .

agreement as long as  $\langle l \rangle - 1 \leq 0.1$ . Hence the turning point can be found by investigating the minimum of  $H_2$ , which gives

$$c_{\text{turn}} \approx -\frac{1}{3} \ln(2\varepsilon), \quad (40)$$

where we have only kept the leading order of  $\varepsilon \ll 1$ . Note that  $c_{\text{turn}}$  indeed diverges as  $\alpha \rightarrow 0$ . When  $c \leq c_{\text{turn}}$ , the mean walk length is well approximated by  $1 + e^{-c}$  which is the result for  $\alpha = 0$  with  $c \geq 1$ .

To put these results into perspective and provide an intuitive explanation of the observed non-monotonic behavior as a function of  $c$ , it is instructive to compare the mean walk length to the density of local fitness maxima. Since the walk is trapped at local maxima, one generally expects an inverse relationship between the two quantities (Weinberger, 1991; Nowak and Krug, 2015). According to (23), the density of local fitness maxima at distance  $d$  from the reference sequence becomes  $p_c^{\text{max}} \approx 1/(\beta L)$  in the limit when  $L, d \rightarrow \infty$  at fixed  $\alpha = d/L$ , where we recall that  $\beta = \alpha e^c + (1 - \alpha)e^{-c}$ . It is straightforward to check that  $p_c^{\text{max}}$  decreases monotonically with increasing  $\alpha$  but displays a maximum as a function of  $c$  for  $\alpha < \frac{1}{2}$ . The maximum is located at  $\tilde{c}_{\text{turn}} = -\frac{1}{2} \ln \varepsilon$  which is similar to (40) and also diverges for  $\alpha \rightarrow 0$ . We may thus conclude that, at least qualitatively, the behavior of the greedy walk length reflects that of the density of local maxima.

## 4. General distribution of the random fitness component

### 4.1. Reformulation of the problem

Up to now, we have presented a detailed analysis of greedy adaptive walks for the case of Gumbel-distributed random fitness components. In this section, we will generalize our findings to arbitrary probability distribution functions  $F(y)$ , focusing on the limit  $L \rightarrow \infty$ . As in Sec. 3.2, the initial genotype from which the walker starts has the Hamming distance  $d_0$  from the reference sequence and we take  $d_0, L \rightarrow \infty$  at fixed

$\alpha = d_0/L$ . Under these conditions the walker takes both uphill and downhill steps. As long as the number of steps taken is much smaller than  $L$ , the walk dynamics can be formulated in terms of the following game:

At each round  $n$  ( $n = 1, 2, \dots$ ), one generates two random variables  $Y_n$  and  $Y_{-n}$ , where  $Y_n$  is drawn from the distribution  $F(y)^{L^\alpha}$  and  $Y_{-n}$  from  $F(y)^{L(1-\alpha)}$ . Then choose the larger one between  $Y_n + c$  and  $Y_{-n} - c$ . Assuming that the larger one is  $Y_{\sigma_n n} + \sigma_n c$  where  $\sigma_n$  can be either 1 or  $-1$ , this number is compared to  $X_{n-1}$ , with  $X_0 = -\infty$ . If  $X_{n-1}$  is larger than  $Y_{\sigma_n n} + \sigma_n c$  the game is over. Otherwise, we set  $X_n = Y_{\sigma_n n} + \sigma_n c$  and go to the next round. Then the mean number of steps in the greedy walk is the same as the mean number of rounds up to the end of the game.

For convenience, we introduce an event

$$E_n(\sigma) = \{Y_{\sigma n} + \sigma c > Y_{-\sigma n} - \sigma c \text{ \& } Y_{\sigma n} + \sigma c > X_{n-1}\}, \quad (41)$$

where  $X_{n-1}$  is defined as above. With this notation, we can write down the probability that the game persists at least up to  $l$  rounds as

$$H_l = \sum_{\{\sigma\}_l} \text{Prob}(E_1(\sigma_1) \cap E_2(\sigma_2) \cap \dots \cap E_l(\sigma_l)), \quad (42)$$

where the summation is over all possible sequences of  $\sigma$ 's of length  $l$ .

For  $\alpha = 1$  all steps are in the uphill direction and (42) reduces to a single term with  $\sigma_1 = \sigma_2 = \dots \sigma_l = 1$ , which can be written as

$$H_l = \text{Prob}(Y_1 + c < Y_2 + 2c < \dots < Y_l + cl), \quad (43)$$

that is, the probability that the sequence of random variables  $Y_n + cn$  is ascendingly ordered. This quantity was studied by Franke et al. (2010) who showed that it is given by (15) when the  $Y_n$ 's are drawn from the Gumbel distribution. To see why this result applies in the present context, we note that the distribution function of the maximum among  $L$  i.i.d. Gumbel random variables is given by

$$F(y)^L = \exp[-Le^{-y}] = \exp[-e^{-(y-\ln L)}] = F(y - \ln L), \quad (44)$$

which is identical to the original distribution up to an overall shift that doesn't affect the ordering probability (43).

#### 4.2. Extreme value classes

In order to analyze the problem for general choices of the distribution function  $F(y)$ , we exploit the fact that  $F(y)^{L^\alpha}$  and  $F(y)^{L(1-\alpha)}$  converge to one of the extreme value distributions when the limit  $L \rightarrow \infty$  is combined with a suitable rescaling of  $y$  (de Haan and Ferreira, 2006). Specifically, we introduce random variables  $Z_k$  such that  $Y_k = a_L Z_k + b_L$ , where  $k$  is an integer,  $a_L$  and  $b_L$  are parameters that depend on  $L$  but not on  $k$ , and  $a_L > 0$ . The parameters  $a_L$  and  $b_L$  have to be chosen such that the distribution of  $Z_k$  has a well defined limit as  $L \rightarrow \infty$ , that is, such that

$$K(z) = \lim_{L \rightarrow \infty} F(a_L z + b_L)^L \quad (45)$$

exists and is non-degenerate.

In terms of the transformed random variables, the event  $E_n(\sigma)$  can be recast as

$$E_n(\sigma) = \{Z_{\sigma n} + \sigma \tilde{c} > Z_{-\sigma n} - \sigma \tilde{c} \text{ \& } Z_{\sigma n} + \sigma \tilde{c} > \tilde{X}_{n-1}\}, \quad (46)$$

where  $\tilde{c} = c/a_L$  and  $\tilde{X}_n = Z_{\sigma_n n}$ . In the following we apply this approach to the three classes of extreme value distributions.

*Gumbel class.* As a representative of the Gumbel class of extreme value theory we choose the Weibull distribution  $F(y) = 1 - e^{-y^\theta}$ . Setting

$$Y_k = (\ln L)^{1/\theta} \left(1 + \frac{Z_k}{\theta \ln L}\right) = (\ln L)^{1/\theta} + \frac{Z_k}{\theta (\ln L)^{1-1/\theta}}, \quad (47)$$

the limit (45) becomes the Gumbel distribution

$$K_G(z) = e^{-e^{-z}} \quad (48)$$

with support  $-\infty < z < \infty$ , as can be seen using the approximation  $y^\theta = \ln L(1+z/(\theta \ln L))^\theta \approx \ln L + z + o(1/\ln L)$ . Accordingly,

$$\tilde{c} = \theta c (\ln L)^{1-1/\theta}. \quad (49)$$

For the case of an exponential distribution ( $\theta = 1$ ) it follows that  $\tilde{c} = c$ , and we conclude that the results derived in Sec. 3 for Gumbel-distributed random fitness components in fact apply asymptotically to *all distributions with exponential tails*. On the other hand, when the tail of the distribution is fatter ( $\theta < 1$ ) or thinner ( $\theta > 1$ ) than exponential,  $\tilde{c}$  asymptotically scales to zero or infinity, respectively, when the limit  $L \rightarrow \infty$  is taken at fixed  $c$ . This implies that greedy adaptive walks on the RMF landscape behave asymptotically like those on an uncorrelated landscape in the first case, their length approaching  $\langle l \rangle = e - 1$ , whereas in the second case the walks move all the way to the reference sequence and  $\langle l \rangle \rightarrow \alpha L$ . Because of the logarithmic dependence of  $\tilde{c}$  on  $L$ , corrections to this asymptotic behavior are however expected to be important, and can be obtained from the results of Sec. 3 by replacing  $c$  with  $\tilde{c}$ .

*Fréchet class.* This class comprises distributions with a power law tail and can be represented by  $F(y) = 1 - y^{-\mu}$  with  $y > 1$  and  $\mu > 0$ . Choosing  $a_L = L^{1/\mu}$  and  $b_L = 0$ , the limit (45) becomes

$$K_F(z) = \lim_{L \rightarrow \infty} \left(1 - \frac{z^{-\mu}}{L}\right)^L = e^{-z^{-\mu}} \quad (50)$$

with the support  $z > 0$ . Accordingly,  $\tilde{c} = c/L^{1/\mu}$ . Assuming that  $c$  remains finite when taking the  $L \rightarrow \infty$  limit,  $\tilde{c}$  approaches zero and the problem becomes identical to the greedy walk on an uncorrelated landscape.

*Weibull class.* Lastly, we consider distributions with bounded support, as represented by the distribution function  $F(y) = 1 - (1 - y)^\nu$  with  $y \in [0, 1]$ . Setting  $a_L = L^{-1/\nu}$  and  $b_L = 1$ , the limiting distribution is

$$K_W(z) = e^{-(z)^\nu} \quad (51)$$



with the support  $z < 0$ . Hence, in this case  $\tilde{c} = cL^{1/\nu}$ . For finite  $c$ ,  $\tilde{c}$  is effectively infinite so that  $H_l = 1$  and  $\langle l \rangle \approx \alpha L$ .

To summarize the results of this section, we have shown that it is only for distributions with exponential tails that the mean greedy walk length displays a non-trivial dependence on  $c$ , and in this case the results of Sec. 3 carry over without modification. In all other cases a non-trivial asymptotic behavior requires that the strength of the fitness gradient  $c$  is scaled with  $L$  in such a way that  $\tilde{c}$  has a finite limit for  $L \rightarrow \infty$ .

For the non-Gumbel extreme value classes characterized by the limiting distributions (50) and (51) a closed-form solution analogous to that obtained in Sec. 3 for the Gumbel class appears to be out of reach, with the exception of the Weibull class with  $\nu = 1$ , where the explicit formula

$$\langle l \rangle = \left( \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} e^{-k(k-1)\tilde{c}/2} \right)^{-1} - 1 \quad (52)$$

can be derived for  $\alpha = 1$  (see Appendix C). In the general case we therefore resort to approximations that are valid for small and large  $\tilde{c}$ , respectively. Apart from their intrinsic interest, these results can be used to compute corrections to the asymptotic walk length when  $c$  and  $L$  are both finite. Throughout we assume a general limiting distribution function  $K(z)$  with the corresponding probability density  $f_K(x) = \frac{dK}{dz}$ .

#### 4.3. Small $\tilde{c}$ approximation

We begin with the case  $\alpha = 1$ , where previous results for the ordering probability (43) can be exploited. Indeed, the results of Franke et al. (2010) imply that

$$H_l = \frac{1}{l!} + \frac{\tilde{c}}{(l-2)!} \int_{-\infty}^{\infty} dx f_K(x)^2 + O(\tilde{c}^2) \quad (53)$$

for  $l \geq 2$  whenever the integral on the right hand side exists (note that  $H_1 = 1$  independent of  $\tilde{c}$ ). Summing over  $l$  it thus follows that

$$\langle l \rangle = e - 1 + e\tilde{c} \int_{-\infty}^{\infty} dx f_K(x)^2 + O(\tilde{c}^2). \quad (54)$$

Although the case of general  $\alpha$  is more complex [as can be seen by comparing the expressions (42) and (43)], the extensive calculations presented in Appendix D and Appendix E yield a simple result which amounts to replacing  $\tilde{c}$  by  $(2\alpha - 1)\tilde{c}$  in (54). Evaluating the integral over  $f_K(x)^2$  for the limiting distributions (48) and (50), we thus obtain

$$\langle l \rangle_G = e - 1 + \frac{2\alpha - 1}{4} e\tilde{c}, \quad (55)$$

$$\langle l \rangle_F = e - 1 + (2\alpha - 1)e\tilde{c}\mu 2^{-2-1/\mu}\Gamma\left(2 + \frac{1}{\mu}\right) \quad (56)$$

to leading order in  $\tilde{c}$ . Note that the result for the Gumbel class is consistent with Eq. (37).

For the Weibull class, the integral on the right hand side of (54) exists only for  $\nu > \frac{1}{2}$ , and a more careful analysis is required to find the leading correction in  $\tilde{c}$ . Detailed calculations

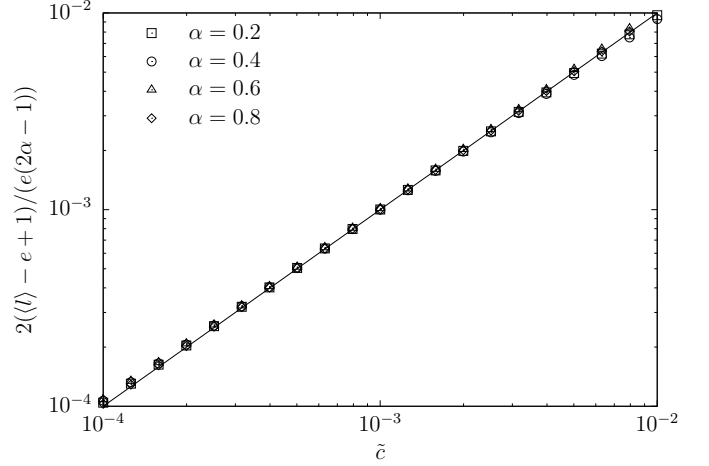


Figure 5: Double logarithmic plot of  $2(\langle l \rangle - e + 1)/(e(2\alpha - 1))$  vs.  $\tilde{c}$  for the Weibull class with  $\nu = 1$  and various values of the scaled initial distance  $\alpha$ . The data sets for different  $\alpha$  collapse into the line  $y = x$  as predicted by Eq. (57).

are found in Appendix E. The final result for the mean greedy walk length in this case reads,

$$\frac{\langle l \rangle_W - e + 1}{e(2\alpha - 1)} = \begin{cases} \nu 2^{-2+1/\nu} \Gamma\left(2 - \frac{1}{\nu}\right) \tilde{c}, & \nu > \frac{1}{2}, \\ -\frac{\tilde{c}}{4} \ln(e^{2\gamma-1} \tilde{c}), & \nu = \frac{1}{2}, \\ \frac{\Gamma(1-2\nu)\Gamma(\nu+1)}{2\Gamma(1-\nu)} \tilde{c}^{2\nu}, & \nu < \frac{1}{2} \end{cases} \quad (57)$$

where  $\gamma \approx 0.5772$  is the Euler-Mascheroni constant.

In Fig. 5, we confirm the validity of Eq. (57) for  $K_W(x) = e^x$  ( $x < 0$ ,  $\nu = 1$ ) by simulations. Note that the result for  $\nu = 1$  and  $\alpha = 1$  can also be obtained by expanding the exact expression (52) up to  $O(\tilde{c})$ . In Fig. 6 we show simulations of Eq. (46) for the Weibull class with various values of  $\nu$  and  $\alpha = 1$ . Each data symbol in Fig. 6 is the result of  $10^{11}$  independent runs. The predicted Eq. (57) is in good agreement with the simulations.

#### 4.4. Large $\tilde{c}$ approximation

When  $\tilde{c}$  is very large, the walker takes uphill steps toward the reference state with probability close to 1 as long as  $\alpha \neq 0$ . (If  $\alpha = 0$  and  $\tilde{c}$  is very large,  $\langle l \rangle \approx 1 + H_2$  as for large negative  $c$  in Sec. 3.2.) In this case, we can neglect the effect of downhill steps and the problem is reduced to the ordering problem studied by Franke et al. (2010) with random variables drawn from a distribution  $K(z)^\alpha$ . If the support of  $K(z)$  is unbounded from the above as in the Gumbel and Fréchet classes, the walker stops at the  $l$ 'th step when  $Z_l$  happens to be larger than  $\tilde{c}$ . Thus  $\langle l \rangle$  can be estimated from the relation  $1 - K(\tilde{c})^\alpha = 1/\langle l \rangle$ . For  $K_G$ , we get  $\langle l \rangle \sim e^{\tilde{c}}/\alpha = \exp[\theta c(\ln L)^{1-1/\theta}]/\alpha$ . For  $K_F$ , we get

$$\langle l \rangle \sim \frac{\tilde{c}^\mu}{\alpha} = \frac{c^\mu}{\alpha L}. \quad (58)$$

If the support of  $K(z)$  is bounded from above but unbounded from below as in the Weibull class, the walker should stop at the

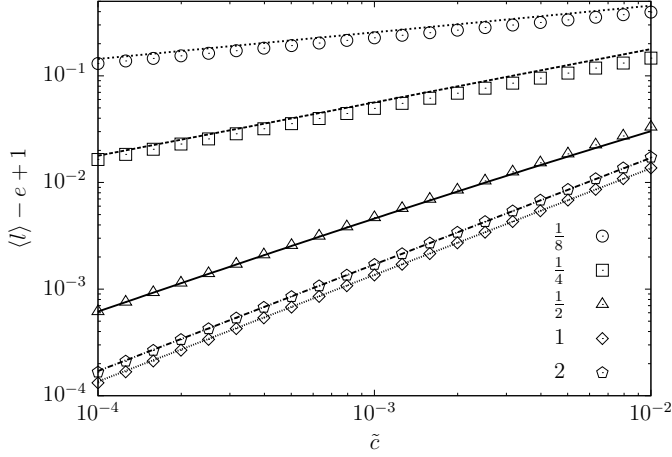


Figure 6: Double logarithmic plot of  $\langle l \rangle - e + 1$  vs.  $\tilde{c}$  for the Weibull class with  $\nu = 2, 1, \frac{1}{2}, \frac{1}{4},$  and  $\frac{1}{8}$ . Here  $\alpha$  is set to 1. The straight lines close to each data set are given by Eq. (57) with the corresponding values of  $\nu$ .

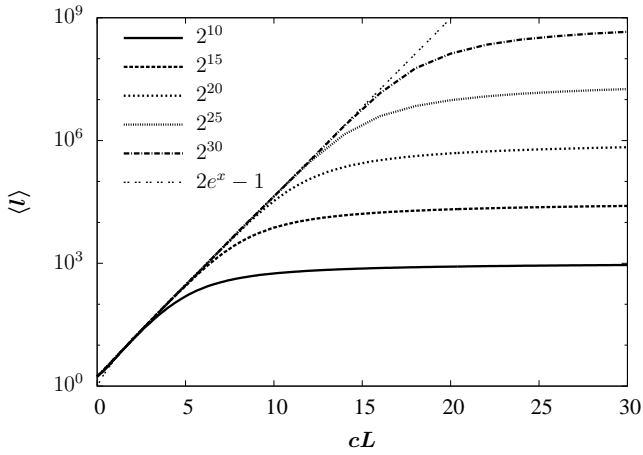


Figure 7: Mean walk length for uniformly distributed random fitness components with sequence length  $L = 2^{10}, 2^{15}, 2^{20}, 2^{25},$  and  $2^{30}$  (from bottom to top) and antipodal starting point ( $\alpha = 1$ ). As predicted by theory, in this case the walk length is a function of  $cL$ . The number of runs is between  $10^3$  (for  $L = 2^{30}$ ) and  $2 \times 10^4$  (for  $L = 2^{10}$ ).

$(l - 1)$ th step when  $Z_l$  happens to be smaller than  $-\tilde{c}$ . Thus  $\langle l \rangle$  can be estimated from  $\langle l \rangle K(-\tilde{c})^\alpha = 1$ , and using the expression for  $K_W$ , we get  $\langle l \rangle \sim e^{\alpha \tilde{c}^\nu} = e^{\alpha c^\nu L}$ . As an example, we present simulation results for the uniform distribution [ $F(x) = x$ ] with  $\alpha = 1$  in Fig. 7. Note that the leading behavior of Eq. (52) for large  $\tilde{c}$  is  $2e^{\tilde{c}} - 1$ , which is consistent with the approximate estimate as well as with the simulation results in Fig. 7.

## 5. Discussion and conclusion

Adaptive walks arise as limiting cases from standard population genetic models and represent an important paradigm in the theory of adaptation that has generated a number of non-trivial and experimentally testable predictions (Orr, 2005; Schoustra et al., 2009; Seetharaman and Jain, 2014). In particular, the greedy adaptive walk considered in the present article is of biological interest for two reasons. First, it can be

viewed as an approximate description of adaptation in a situation where the supply of single beneficial mutations is high, such that all mutants are generated simultaneously and the mutation of largest effect takes over by selection. Second, the greedy search strategy is arguably one that locates local fitness maxima in the smallest possible number of steps. Greedy walks therefore provide important insights into the geometry of high-dimensional random fitness landscapes, where, as shown by Orr (2003) for the uncorrelated case, fitness peaks are found within 2 mutational steps on average.

Here we have generalized the analysis of Orr (2003) to the class of RMF models, where a fitness gradient of strength  $c$  is introduced to smoothen the landscape and to induce correlations between genotypes. Fitness correlations are generally expected to increase the length of adaptive walks, and we show that this is true in most but not all situations.

Importantly, we find that the effect of the fitness gradient on the length of greedy walks depends crucially on the tail properties of the distribution underlying the random fitness component of the RMF landscape, which can be classified in terms of extreme value theory (EVT). The results of our analysis in Sec. 4 imply that greedy walks on the RMF landscape are asymptotically as short as in the uncorrelated case when the distribution of the random fitness contribution is heavy tailed (Fréchet or Gumbel with tail fatter than exponential) but become very long, with length equal to the distance to the reference sequence, when the distribution is light tailed (Weibull or Gumbel with tail thinner than exponential). Analogous results that single out fitness distributions with exponential tails with regard to structural properties of the RMF landscape (such as the number of local fitness peaks) and the length of random adaptive walks on this landscape have been reported previously (Neidhart et al., 2014; Park et al., 2015).

The prime representative of exponentially-tailed fitness distributions in the Gumbel class of EVT is the Gumbel distribution itself. For this case detailed results for the distribution of the greedy walk length were obtained in Sec. 3, for finite  $L$  and antipodal starting point as well as for arbitrary starting points and  $L \rightarrow \infty$ . Perhaps the most surprising result of our analysis is the finding that the mean walk length depends non-monotonically on the strength of the fitness gradient  $c$ , when the walk starts closer than at distance  $L/2$  from the reference sequence (i.e., the scaled distance is  $\alpha < \frac{1}{2}$ ). This behavior was first observed numerically by Neidhart et al. (2014), and we have argued that it can be related to a similar non-monotonicity in the local density of fitness peaks.

Although the analysis in Sec. 3.2 is restricted to Gumbel-distributed fitnesses, the fact that the leading order correction to the uncorrelated walk length derived in Sec. 4 is universally proportional to  $2\alpha - 1$ , and hence changes sign at  $\alpha = \frac{1}{2}$ , indicates that the phenomenon is robust and does not depend on the distribution of the random fitness component. Notably, within the EVT of adaptation it is usually assumed that the fitness of the wild type is high in absolute terms (Gillespie, 1984; Orr, 2002, 2005). In the context of the RMF model this implies that the adaptive walk starts rather close to the reference sequence, that is, at small  $\alpha$ , where the minimum in the walk length as a

function of  $c$  is particularly pronounced (see Sec. 3.2).

The existence of this minimum appears to contradict Orr's conjecture that the  $c = 0$  value  $\langle l \rangle = e - 1$  constitutes a general lower bound on the length of adaptive walks (Orr, 2003). However, in formulating his conjecture Orr demanded that the walk starts at a randomly chosen point in sequence space, which implies that our result in Eq. (37) should be averaged over  $\alpha$ . Since the probability of choosing  $\alpha$  is symmetric under the transformation  $\alpha \mapsto 1 - \alpha$ , the average of  $2\alpha - 1$  is zero and that of  $\alpha(1 - \alpha)$  is positive. It follows that the averaged walk length cannot be smaller than  $e - 1$ . Similarly, in Rosenberg's refinement of Orr's conjecture it is postulated that the fitness values in the landscape are identically distributed, and that the fitness correlations between neighboring genotypes are positive (Rosenberg, 2005, Sec. 5). Whereas the latter statement applies to the RMF model (Neidhart et al., 2014), the former does not. We conclude, therefore, that the seeming violation of the conjecture of Orr (2003) must be attributed to the anisotropy of the RMF landscape.

An important aspect of adaptation that we have not addressed in this work concerns the fitness level reached by the population at the end of the adaptive walk. In a recent comparative study of different types of adaptive walks on Kauffman's NK-landscape, it was found that greedy walks reach higher fitness levels than random adaptive walks on correlated landscapes, but the ranking among the walk types may change in the presence of correlations (Nowak and Krug, 2015). The results presented here suggest that a detailed analysis of the interplay between fitness correlations and the efficiency of different modes of adaptation may be feasible within the framework of the RMF model, and we hope to report results along these lines in the future.

## Acknowledgments

S-CP acknowledges the support by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant No. 2014R1A1A2058694); and by The Catholic University of Korea, Research Fund, 2015. JN and JK acknowledge support by Deutsche Forschungsgemeinschaft within SFB-TR12, SPP 1590 and BCGS. Computations were performed on the Cheops cluster at RRZK, Universität zu Köln.

## Appendix A. Simulation Method

Since we only need the largest value among a certain number of i.i.d. random variables with a known distribution function, only two random numbers are necessary to check if the walker can take a further step (see also Sec. 4.1). To be concrete, let us assume that the walker is at  $C$  with fitness  $-dc + \xi$  where  $d$  is the Hamming distance of  $C$  from the reference sequence. Since the cumulative distribution of the largest random number among  $k$  variables is  $F(x)^k$ ,  $x$  can be generated by  $x = F^{-1}(y^{1/k})$ , where  $y$  is a uniformly distributed random number ( $0 < y < 1$ ). For the Gumbel distribution,  $x = \ln k - \ln(-\ln y)$  and for the uniform distribution,  $x = y^{1/k}$ .

If  $x_1$  ( $x_2$ ) is the largest random number among the uphill (downhill) neighbors and if either  $x_1 + c$  or  $x_2 - c$  is larger than  $\xi$ , the walker takes one further step. The actual direction will be determined by checking whether  $x_1 + 2c > x_2$  or not. If  $\xi$  is the largest, the walker stops.

If  $k$  is very large, we sometimes use the following approximation

$$y^{1/k} = e^{\ln y/k} \approx 1 + \frac{\ln y}{k} \left( 1 + \frac{\ln y}{2k} \right). \quad (\text{A.1})$$

As a rule of thumb, for  $k \geq 50\,000$ , the above approximation gives a more accurate value than the direct power calculation when we perform numerics with double precision ( $\sim 10^{-15}$ ). Note that when  $k$  is very large, it is better to use  $1 - x$  when deciding the fate of the walk, otherwise there could be round-off errors which give  $x = 1$ .

## Appendix B. Derivation of Eq. (36)

We first derive Eq. (33). The integral over  $y_1$  is readily calculated as

$$L\beta \int_{-\infty}^{y_2 + \sigma_2 c} \exp(-y_1 - \sigma_1 c - L\beta e^{-\sigma_1 c} e^{-y_1}) dy_1 = \exp(-L\beta e^{-(\sigma_1 + \sigma_2)c} e^{-y_2}). \quad (\text{B.1})$$

Using the above equation, we can calculate the integral over  $y_2$  as

$$L\beta \int_{-\infty}^{y_3 + \sigma_3 c} \exp(-y_2 - \sigma_2 c - L\beta e^{-\sigma_2 c} (1 + e^{-\sigma_1 c}) e^{-y_2}) dy_2 = \frac{1}{1 + e^{-\sigma_1 c}} \exp(-L\beta e^{-\sigma_3 c} (e^{-\sigma_1 c} + e^{-(\sigma_1 + \sigma_2)c}) e^{-y_3}), \quad (\text{B.2})$$

from which one can easily guess and prove that

$$(L\beta)^{n-1} \prod_{k=n}^2 \int_{-\infty}^{y_k + \sigma_k c} Q(y_{k-1} + \sigma_{k-1} c) dy_{k-1} = \prod_{k=1}^{n-1} \frac{1}{1 + \sum_{m=1}^{k-1} e^{-cM_m}} \exp\left(-L\beta e^{-\sigma_n c} \sum_{m=1}^{n-1} e^{-cM_m} e^{-y_n}\right). \quad (\text{B.3})$$

The final integral over  $y_l$  gives Eq. (33).

For small  $c$ , we first expand the denominator in Eq. (33) up to  $O(c^2)$ , which is

$$1 + \sum_{m=1}^{k-1} e^{-cM_m} = k \left( 1 - \frac{c}{k} \sum_{m=1}^{k-1} M_m + \frac{c^2}{2k} \sum_{m=1}^{k-1} M_m^2 \right). \quad (\text{B.4})$$

Then we expand the terms in  $H_l$  up to  $O(c^2)$  to get

$$l!H_l = \sum_{\{\sigma\}} S(\{\sigma\}) \left( 1 + c\gamma_1 + \frac{c^2}{2} (\gamma_2 + \gamma_3 - \gamma_4) \right), \quad (\text{B.5})$$

where  $S(\{\sigma\}) = \prod_{k=1}^l s_{\sigma_k}$  and

$$\gamma_1 = \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \sum_{n=1}^m \sigma_n, \quad (\text{B.6})$$

$$\gamma_2 = \sum_{k=1}^l \frac{1}{k^2} \left( \sum_{m=1}^{k-1} \sum_{n=1}^m \sigma_n \right)^2, \quad (\text{B.7})$$

$$\gamma_3 = \left( \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \sum_{n=1}^m \sigma_n \right)^2, \quad (\text{B.8})$$

$$\gamma_4 = \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \left( \sum_{n=1}^m \sigma_n \right)^2. \quad (\text{B.9})$$

The summations over  $\sigma$  in Eq. (B.5) have two forms

$$\sum_{\{\sigma\}} S(\{\sigma\}) \sigma_m = (s_1 - s_{-1}) \prod_{k \neq m} \sum_{\sigma_k} s_{\sigma_k} = \delta, \quad (\text{B.10})$$

$$\sum_{\{\sigma\}} S(\{\sigma\}) \sigma_m \sigma_n = \delta^2 + \delta_{mn}(1 - \delta^2) \quad (\text{B.11})$$

where  $\delta = s_1 - s_{-1}$  and we have used  $s_1 + s_{-1} = 1$  and  $\sigma_n^2 = 1$ . Thus, we get

$$\begin{aligned} \gamma_1 &= \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \sum_{n=1}^m \sum_{\sigma} S(\{\sigma\}) \sigma_n = \delta \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} m \\ &= \frac{l(l-1)}{4} \delta, \end{aligned} \quad (\text{B.12})$$

$$\begin{aligned} \gamma_2 &= \sum_{k=1}^l \frac{1}{k^2} \sum_{m=1}^{k-1} \sum_{n=1}^m \sum_{r=1}^{k-1} \sum_{s=1}^r \sum_{\{\sigma\}} S(\{\sigma\}) \sigma_s \sigma_n \\ &= \sum_{k=1}^l \frac{1}{k^2} \sum_{m=1}^{k-1} \sum_{n=1}^m \sum_{r=1}^{k-1} \sum_{s=1}^r (\delta^2 + \delta_{sn}(1 - \delta^2)) \\ &= \delta^2 \sum_{k=1}^l \frac{(k-1)^2}{4} + (1 - \delta^2) \sum_{k=1}^l \frac{1}{k^2} \sum_{m=1}^{k-1} \sum_{r=1}^{k-1} \min(m, r) \\ &= \delta^2 \frac{l(l-1)(2l-1)}{24} + (1 - \delta^2) \sum_{k=1}^l \frac{(k-1)(2k-1)}{6k} \\ &= \frac{\delta^2}{24} l(l-1)(2l-1) + \frac{1 - \delta^2}{6} (l^2 - 2l + \text{Har}[l]), \end{aligned} \quad (\text{B.13})$$

where  $\text{Har}[l] = \sum_{k=1}^l k^{-1}$ , and

$$\begin{aligned} \gamma_3 &= \sum_{k_1=1}^l \sum_{k_2=1}^l \frac{1}{k_1 k_2} \sum_{m=1}^{k_1-1} \sum_{n=1}^m \sum_{r=1}^{k_2-1} \sum_{s=1}^r (\delta^2 + \delta_{sn}(1 - \delta^2)) \\ &= \delta^2 \left( \sum_{k=1}^l \frac{k-1}{2} \right)^2 + (1 - \delta^2) \sum_{k_1=1}^l \sum_{k_2=1}^l \frac{1}{k_1 k_2} \sum_{m=1}^{k_1-1} \sum_{r=1}^{k_2-1} \min(m, r) \\ &= \delta^2 \frac{(l(l-1))^2}{16} + (1 - \delta^2) \sum_{k_1=1}^l \sum_{k_2=1}^l \frac{(k-1)(3K - k - 1)}{6K}, \end{aligned} \quad (\text{B.14})$$

where  $K = \max(k_1, k_2)$  and  $k = \min(k_1, k_2)$ . The summation in the last line of the above equation is

$$\begin{aligned} &\sum_{k=1}^l \frac{(k-1)(2k-1)}{6k} + 2 \sum_{K=2}^l \sum_{k=1}^{K-1} \frac{(k-1)(3K - k - 1)}{6K} \\ &= \frac{l(14l^2 - 33l + 37)}{108} - \frac{1}{6} \text{Har}[l]. \end{aligned} \quad (\text{B.15})$$

And finally

$$\begin{aligned} \gamma_4 &= \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \sum_{n=1}^m \sum_{r=1}^m \sum_{\{\sigma\}} S(\{\sigma\}) \sigma_n \sigma_r \\ &= \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} \sum_{n=1}^m \sum_{r=1}^m (\delta^2 + \delta_{nr}(1 - \delta^2)) \\ &= \sum_{k=1}^l \frac{1}{k} \sum_{m=1}^{k-1} (\delta^2 m^2 + m(1 - \delta^2)) \\ &= \frac{l(l-1)}{4} + \delta^2 \frac{l(l-1)(l-2)}{9} \end{aligned} \quad (\text{B.16})$$

Combining these results, we arrive at Eq. (36).

### Appendix C. Derivation of Eq. (52)

This appendix calculates the mean walk length for the case of  $K(x) = e^x$  ( $x < 0$ ) with  $\alpha = 1$ . For brevity, we will denote the mean walk length by  $\ell$  and we drop the tilde in  $\tilde{c}$  in this appendix. For completeness, we write the probability  $H_l$  of taking at least  $l$  steps

$$H_l = \int_{-\infty}^0 j_l(x) dy, \quad (\text{C.1})$$

where  $j_l(x)$  satisfies the recursion relation

$$j_l(x) = e^x \int_{-\infty}^{x+c} j_{l-1}(x), \quad (\text{C.2})$$

with  $j_1(x) = e^x$ . Since the support of  $j_l(x)$  is  $x < 0$ , Eq. (C.2) for  $-c < x$  should be interpreted as

$$j_l(x) = e^x H_{l-1}. \quad (\text{C.3})$$

Let

$$\varphi(x) = \int_{-\infty}^x \sum_{l=1}^{\infty} j_l(y) dy. \quad (\text{C.4})$$

which is related to the mean walk distance by  $\ell = \sum_l H_l = \varphi(0)$ . By summing both sides of Eq. (C.2) from  $l = 2$  to infinity, we get the difference-differential equation

$$\frac{d\varphi(x)}{dx} = e^x (1 + \varphi(x + c)), \quad (\text{C.5})$$

where  $\varphi(x)$  with  $x > 0$  should be interpreted as  $\ell$ . Thus, for  $x > -c$ , we have

$$\varphi(x) - \varphi(-c) = (1 + \ell)(e^x - e^{-c}). \quad (\text{C.6})$$



Using  $\varphi(0) = \ell$ , we get  $\varphi(-c) = (1 + \ell)e^{-c} - 1$  which, in turn, gives, for  $x > -c$ ,

$$\varphi(x) = (1 + \ell)e^x - 1. \quad (\text{C.7})$$

Having determined  $\varphi(x)$  for  $x > -c$ , we can find  $\varphi(x)$  for  $-2c < x < -c$  and so on. After a few attempts, we make an ansatz, for  $-nc < x < -(n-1)c$ ,

$$\varphi(x) = (1 + \ell) \sum_{k=0}^n \frac{a_{n-k}}{k!} \exp(kx + k(k-1)c/2) - 1, \quad (\text{C.8})$$

which satisfies Eq. (C.5). From the continuity of  $\varphi$  at  $x = -nc$ , that is,  $\varphi(-nc + 0) = \varphi(-nc - 0)$ , we get a recursion relation for the  $a_n$  as

$$a_{n+1} = e^{-n(n+1)c/2} \sum_{k=0}^n a_k \left( \frac{e^{k(k+1)c/2}}{(n-k)!} - \frac{e^{k(k-1)c/2}}{(n-k+1)!} \right), \quad (\text{C.9})$$

which is identical to

$$(a_{n+1} - a_n) e^{n(n+1)c/2} = - \sum_{k=0}^{n-1} \frac{1}{(n-k)!} (a_{k+1} - a_k) e^{k(k+1)c/2} - \frac{1}{(n+1)!}, \quad (\text{C.10})$$

with  $a_0 = 1$  and  $a_1 = 0$ . If we define  $d_k = (k+1)!(a_{k+1} - a_k)e^{k(k+1)c/2}$ , we get

$$d_n = - \sum_{k=0}^{n-1} \binom{n+1}{k+1} d_k - 1 \quad (\text{C.11})$$

or

$$\sum_{k=0}^n \binom{n+1}{k+1} d_k = -1. \quad (\text{C.12})$$

Since  $d_0 = -1$  and  $\sum_{k=0}^n \binom{n+1}{k+1} (-1)^{k+1} = -1$ , we conclude that  $d_k = (-1)^{k+1}$ . That is, we get the recursion

$$a_{n+1} - a_n = \frac{(-1)^{n+1}}{(n+1)!} e^{-n(n+1)c/2} \quad (\text{C.13})$$

which is solved by

$$a_n = \sum_{k=0}^n \frac{(-1)^k}{k!} e^{-k(k-1)c/2}. \quad (\text{C.14})$$

The mean walk length  $\ell$  is determined by the boundary condition  $\varphi(-\infty) = 0$ . Since  $e^{kx}$  decays exponentially to zero unless  $k = 0$ , this condition becomes

$$0 = \varphi(-\infty) = -1 + (\ell + 1) \lim_{n \rightarrow \infty} a_n \quad (\text{C.15})$$

which gives Eq. (52).

## Appendix D. Small $\tilde{c}$ behavior : formal derivation

In this appendix, we present a formal derivation of the small  $\tilde{c}$  behavior reported in Sec. 4.3. In analogy to (28) we first introduce

$$q_K(z, \sigma, \tilde{c}) = \omega_\sigma f_K(z) \left[ \frac{K(z + 2\sigma\tilde{c})}{K(z)} \right]^{1-\omega_\sigma}, \quad (\text{D.1})$$

where  $\omega_{+1} = \alpha$ ,  $\omega_{-1} = 1 - \alpha$ , and  $f_K(z) = \frac{dK(z)}{dz}$ . We also introduce  $j_l$  iteratively as

$$j_l(x, \{\sigma\}_l) = q_K(x, \sigma_l, \tilde{c}) \int_{-\infty}^{x+\sigma_l\tilde{c}} j_{l-1}(y, \{\sigma\}_{l-1}) dy, \quad (\text{D.2})$$

with  $j_1(x, \{\sigma\}_1) = q_K(x, \sigma_1, \tilde{c})$ . Using  $j_l$ , we can write

$$\langle l \rangle = \sum_{l=1}^{\infty} \sum_{\{\sigma\}_l} \int_{-\infty}^{\infty} j_l(x, \{\sigma\}_l) dx, \quad (\text{D.3})$$

where  $\sum_{\{\sigma\}_l}$  stands for the summation over all possible  $\sigma_i$ 's for  $i = 1, \dots, l$ .

Since the mean walk distance for  $\tilde{c} = 0$  is  $e - 1$  for any distribution,  $\langle l \rangle$  should take the form

$$\langle l \rangle = e - 1 + \lambda(\tilde{c}) \quad (\text{D.4})$$

with the property that  $\lambda(\tilde{c}) \rightarrow 0$  as  $\tilde{c} \rightarrow 0$ . In this appendix, we find the leading behavior of  $\lambda(\tilde{c})$  for small  $\tilde{c} \ll 1$ . At first, we decompose  $j_l(x, \{\sigma\}_l) = j_l^{(0)}(x, \{\sigma\}_l) + g_l(x, \{\sigma\}_l, \tilde{c})$ , where

$$\begin{aligned} j_l^{(0)}(x, \{\sigma\}_l) &= q_K^0(x, \sigma_l) \frac{A(\{\sigma\}_{l-1})}{(l-1)!} K(x)^{l-1} \\ &= A(\{\sigma\}_l) \frac{1}{l!} \frac{d}{dx} [K(x)]^l, \end{aligned} \quad (\text{D.5})$$

with

$$q_K^0(z, \sigma) \equiv q_K(z, \sigma, \tilde{c} = 0) = \omega_\sigma f_K(z), \quad (\text{D.6})$$

$$A(\{\sigma\}_l) \equiv \prod_{i=1}^l \omega_{\sigma_i}, \quad A(\{\sigma\}_0) \equiv 1, \quad (\text{D.7})$$

and  $g_l$  satisfies the recursion relation

$$\begin{aligned} g_l(x, \{\sigma\}_l, \tilde{c}) &= k_{l,0}(x, \{\sigma\}_l, \tilde{c}) \\ &+ q_K(x, \sigma_l, \tilde{c}) \int_{-\infty}^{x+\sigma_l\tilde{c}} g_{l-1}(y, \{\sigma\}_{l-1}, \tilde{c}) dy, \end{aligned} \quad (\text{D.8})$$

with

$$\begin{aligned} k_{l,0}(x, \{\sigma\}_l, \tilde{c}) &\equiv \frac{A(\{\sigma\}_{l-1})}{(l-1)!} \left[ q_K(x, \sigma_l, \tilde{c}) K(x + \sigma_l\tilde{c})^{l-1} \right. \\ &\quad \left. - q_K^0(x, \sigma_l) K(x)^{l-1} \right]. \end{aligned} \quad (\text{D.9})$$

Note that  $j_l^{(0)}$  is the solution of Eq. (D.2) for  $\tilde{c} = 0$ . Defining  $k_{l,m}$  recursively as ( $m \geq 1$ )

$$\frac{k_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c})}{q_K(x, \sigma_{l+m}, \tilde{c})} = \int_{-\infty}^{x+\sigma_{l+m}\tilde{c}} k_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy, \quad (\text{D.10})$$

we can formally write

$$g_l(x, \{\sigma\}_l, \tilde{c}) = \sum_{m=0}^{l-1} k_{l-m,m}(x, \{\sigma\}_l, \tilde{c}). \quad (\text{D.11})$$

If we define

$$a_{l,m}(\{\sigma\}_{l+m}, \tilde{c}) \equiv \int_{-\infty}^{\infty} k_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}) dx, \quad (\text{D.12})$$

we can write

$$\lambda(\tilde{c}) = \sum_{l=1}^{\infty} \sum_{m=0}^{\infty} \sum_{\{\sigma\}} a_{l,m}(\{\sigma\}_{l+m}, \tilde{c}), \quad (\text{D.13})$$

where  $\lambda(\tilde{c})$  is defined in Eq. (D.4). Since, for any  $x$  and  $\sigma$ ,

$$\left| \int_{-\infty}^{x+\sigma\tilde{c}} k_{l,m}(y, \{\sigma\}_{l+m}, \tilde{c}) dy \right| \leq \left| \int_{-\infty}^{\infty} k_{l,m}(y, \{\sigma\}_{l+m}, \tilde{c}) dy \right|, \quad (\text{D.14})$$

we get an inequality for any  $l$  and  $m$  such as

$$|a_{l,m}| \leq \int_{-\infty}^{\infty} q_K(x, \sigma_{l+m}, \tilde{c}) dx |a_{l,m-1}| = |a_{l,m-1}| \leq |a_{l,0}|, \quad (\text{D.15})$$

which shows that  $a_{l,m} = O(a_{l,0})$ .

To extract the leading behavior of  $\lambda(\tilde{c})$ , we consider the derivative of  $a_{l,m}$  with respect to  $\tilde{c}$ . That is, we consider

$$\begin{aligned} b_{l,m}(\{\sigma\}_{l+m}, \tilde{c}) &\equiv \frac{\partial}{\partial \tilde{c}} a_{l,m}(\{\sigma\}_{l+m}, \tilde{c}) \\ &= \int_{-\infty}^{\infty} \tilde{k}_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}) dx, \end{aligned} \quad (\text{D.16})$$

where  $\tilde{k}_{l,m}$  is defined as

$$\tilde{k}_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}) \equiv \frac{\partial}{\partial \tilde{c}} k_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}). \quad (\text{D.17})$$

Notice that

$$\frac{d\lambda(\tilde{c})}{d\tilde{c}} = \sum_{l=1}^{\infty} \sum_{m=0}^{\infty} \sum_{\{\sigma\}} b_{l,m}(\{\sigma\}_{l+m}, \tilde{c}). \quad (\text{D.18})$$

If  $b_{l,m} = O(\tilde{c}^\eta)$  with  $\eta \leq 1$  for all  $m$ , we can write  $\tilde{k}_{l,m}$  as the sum of  $\kappa_{l,m}$  and  $R_{l,m}$  which have the property that

$$\begin{aligned} \int_{-\infty}^{\infty} \kappa_{l,m}(y, \{\sigma\}_{l+m}, \tilde{c}) dy &= O(\tilde{c}^\eta), \\ \int_{-\infty}^{\infty} R_{l,m}(y, \{\sigma\}_{l+m}, \tilde{c}) dy &= o(\tilde{c}^\eta). \end{aligned} \quad (\text{D.19})$$

Now we will find a recursion relation for  $\kappa_{l,m}$  from the exact

relation,  $\tilde{k}_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}) = \sum_{i=1}^6 I_i$ , where

$$\begin{aligned} I_1 &= q_K^0(x, \sigma_{l+m}) \int_{-\infty}^x \kappa_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy, \\ I_2 &= q_K^0(x, \sigma_{l+m}) \int_{-\infty}^x R_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy, \\ I_3 &= q_K^0(x, \sigma_{l+m}) \int_x^{x+\sigma_{l+m}\tilde{c}} \tilde{k}_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy \\ I_4 &= q_K^1(x, \sigma_{l+m}, \tilde{c}) \int_{-\infty}^{x+\sigma_{l+m}\tilde{c}} \kappa_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy, \\ I_5 &= q_K(x, \sigma_{l+m}, \tilde{c}) \kappa_{l,m-1}(x + \sigma_{l+m}\tilde{c}, \{\sigma\}_{l+m-1}, \tilde{c}), \\ I_6 &= [q_K - q_K^0] \int_{-\infty}^{x+\sigma_{l+m}\tilde{c}} \tilde{k}_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy, \end{aligned}$$

and

$$\begin{aligned} q_K^1(x, \sigma, \tilde{c}) &= \frac{\partial}{\partial \tilde{c}} q_K(x, \sigma, \tilde{c}) \\ &= 2\sigma\alpha(1-\alpha) \frac{f_K(x)f_K(x+2\sigma\tilde{c})}{K(x+2\sigma\tilde{c})} \left( \frac{K(x+2\sigma\tilde{c})}{K(x)} \right)^{1-\omega_\sigma}. \end{aligned} \quad (\text{D.20})$$

Since  $I_3, I_4, I_5, I_6$  are zero if  $\tilde{c} = 0$  (note that  $\kappa_{l,m}$  is zero if  $\tilde{c} = 0$ ), the integrals over these functions are  $o(1)$  and they contribute to the remainder terms  $R_{l,m}$ . Since the integral of  $I_2$  should be  $o(\tilde{c}^\eta)$ , we find  $R_{l,m} = \sum_{i=2}^5 I_i$  and

$$\kappa_{l,m}(x, \{\sigma\}_{l+m}, \tilde{c}) = q_K^0(x, \sigma_{l+m}) \int_{-\infty}^x \kappa_{l,m-1}(y, \{\sigma\}_{l+m-1}, \tilde{c}) dy. \quad (\text{D.21})$$

In fact, the above consideration reveals that if we choose  $R_{l,0}$  such that the integral of  $R_{l,0}(y)$  is  $o(1)$ , the analysis of  $\kappa_{l,m}$  should give all terms up to  $O(1)$ .

If we define

$$\phi(x) = \sum_{l=1}^{\infty} \sum_{m=0}^{\infty} \sum_{\{\sigma\}_{l+m}} \int_{-\infty}^x \kappa_{l,m}(y, \{\sigma\}_{l+m}, \tilde{c}) dy, \quad (\text{D.22})$$

we can write a differential equation for  $\phi(x)$  such as

$$\frac{d\phi(x)}{dx} = f_K(x)\phi(x) + \chi(x, \tilde{c}) \quad (\text{D.23})$$

where

$$\chi(x, \tilde{c}) = \sum_{l=1}^{\infty} \sum_{\{\sigma\}_l} \kappa_{l,0}(x, \{\sigma\}_l, \tilde{c}), \quad (\text{D.24})$$

and we have used  $\sum_{\sigma} q_K^0(x, \sigma) = f_K(x)$ . The solution of Eq. (D.23) is

$$\phi(x) = e^{K(x)} \int_{-\infty}^x e^{-K(y)} \chi(y, \tilde{c}) dy, \quad (\text{D.25})$$

which is related to  $\lambda(\tilde{c})$  as

$$\frac{d\lambda(\tilde{c})}{d\tilde{c}} \approx \lim_{x \rightarrow \infty} \phi(x) = e \int_{-\infty}^{\infty} e^{-K(y)} \chi(y, \tilde{c}) dy = e\psi(\tilde{c}), \quad (\text{D.26})$$

and, in turn,

$$\langle l \rangle \approx e - 1 + e \int_0^{\tilde{c}} \psi(x) dx, \quad (\text{D.27})$$

where the definition of  $\psi$  is clear from the context.

If we choose  $\kappa_{l,0} = \tilde{k}_{l,0}$  and  $R_{l,0} = 0$ ,  $\phi(x)$  can be used to find all terms up to  $O(1)$ . With this choice, we get

$$\begin{aligned} \chi(x, \tilde{c}) &= \frac{\partial}{\partial \tilde{c}} \left[ \sum_{l=1}^{\infty} \sum_{\{\sigma\}_l} k_{l,0}(x, \{\sigma\}_l, \tilde{c}) \right] \\ &= \sum_{\sigma} \left[ q_K^1(x, \sigma, \tilde{c}) + \sigma q_K(x, \sigma, \tilde{c}) f_K(x + \sigma \tilde{c}) \right] e^{K(x + \sigma \tilde{c})}. \end{aligned} \quad (\text{D.28})$$

### Appendix E. Small $\tilde{c}$ behavior : explicit formulae

This appendix is a continuation of Appendix D and presents the explicit small  $\tilde{c}$  behavior for the various classes. We first assume that  $\psi(0)$  is non-zero. If this is true, the mean walk distance becomes

$$\langle l \rangle = e - 1 + e \tilde{c} \psi(0) + o(\tilde{c}). \quad (\text{E.1})$$

Since  $\sum_{\sigma} q_K^1(x, \sigma, \tilde{c} = 0) = 0$  and, in turn,  $\chi(x, 0) = (2\alpha - 1)f_K(x)^2 \exp[K(x)]$ , we get

$$\psi(0) = e(2\alpha - 1) \int_{-\infty}^{\infty} f_K(x)^2 dx, \quad (\text{E.2})$$

and, accordingly,

$$\langle l \rangle = e - 1 + e \tilde{c} (2\alpha - 1) \int_{-\infty}^{\infty} f_K(x)^2 dx + o(\tilde{c}), \quad (\text{E.3})$$

as long as the integral is finite. As advertised, this generalizes Eq. (54) to  $\alpha < 1$ . For the Gumbel and Fréchet classes and for the Weibull class with  $\nu > \frac{1}{2}$ , the integral becomes

$$\int_{-\infty}^{\infty} e^{-2x-2e^{-x}} dx = \frac{1}{4}, \quad (\text{E.4})$$

$$\int_0^{\infty} \mu^2 x^{-2(\mu+1)} e^{-2x^{-\mu}} dx = \mu 2^{-2-1/\mu} \Gamma(2 + \mu^{-1}), \quad (\text{E.5})$$

$$\int_{-\infty}^0 \nu^2 (-x)^{2(\nu-1)} e^{-2(-x)^{\nu}} dx = \nu 2^{-2+1/\nu} \Gamma(2 - \nu^{-1}). \quad (\text{E.6})$$

For the Weibull class with  $\nu \leq \frac{1}{2}$ , Eq. (E.6) is not applicable and we have to be more careful to find the small  $\tilde{c}$  behavior of  $\psi(\tilde{c})$  for this case. To this end, we first write

$$\psi(\tilde{c}) = \sum_{\sigma} [N_1(\sigma, \tilde{c}) + N_2(\sigma, \tilde{c})], \quad (\text{E.7})$$

where

$$N_1(\sigma, \tilde{c}) = \int_{-\infty}^{\infty} q_K^1(x, \sigma, \tilde{c}) e^{K(x + \sigma \tilde{c}) - K(x)} dx, \quad (\text{E.8})$$

$$N_2(\sigma, \tilde{c}) = \sigma \int_{-\infty}^{\infty} q_K(x, \sigma, \tilde{c}) f(x + \sigma \tilde{c}) e^{K(x + \sigma \tilde{c}) - K(x)} dx. \quad (\text{E.9})$$

We begin with the analysis of  $N_1(\sigma, \tilde{c})$ . Since the support of  $f_K(x)$  in question is  $x < 0$ , we write  $N_1(\sigma, \tilde{c})$  as

$$\begin{aligned} \frac{N_1(\sigma, \tilde{c})}{2\sigma\nu^2\alpha(1-\alpha)} &= \int_{-\infty}^{-2u(\sigma)\tilde{c}} dx [x(x + 2\sigma\tilde{c})]^{y-1} e^{-\zeta_{11}(-x, \sigma, \tilde{c})} \\ &= \int_0^{\infty} dy [y(y + 2\tilde{c})]^{y-1} e^{-\zeta_{11}(y+2u(\sigma)\tilde{c}, \sigma, \tilde{c})}, \end{aligned} \quad (\text{E.10})$$

where  $u(\sigma) = \max(0, \sigma)$ , we have changed variables to  $y = -x - 2u(\sigma)\tilde{c}$  to get the second line, and

$$\zeta_{11}(z, \sigma, \tilde{c}) = z^{\nu} + (1 - \omega_{\sigma}) [(z - 2\sigma\tilde{c})^{\nu} - z^{\nu}] + e^{-(z - \sigma\tilde{c})^{\nu}} - e^{-z^{\nu}}. \quad (\text{E.11})$$

If we again change variables to  $z = y/\tilde{c}$ , the above integral becomes

$$\frac{N_1(\sigma, \tilde{c})}{2\sigma\nu^2\alpha(1-\alpha)} = \tilde{c}^{2\nu-1} \int_0^{\infty} [z(z + 2)]^{y-1} e^{-\zeta_{12}(z, \sigma, \tilde{c})} dy, \quad (\text{E.12})$$

where  $\zeta_{12}(z, \sigma, \tilde{c}) = \zeta_{11}(\tilde{c}z + 2\tilde{c}u(\sigma), \sigma, \tilde{c})$ , which is zero if  $\tilde{c} = 0$ . Thus, the leading behavior of  $N_1$  is

$$\frac{N_1(\sigma, \tilde{c})}{2\nu^2\alpha(1-\alpha)} = \sigma \tilde{c}^{2\nu-1} \int_0^{\infty} [y(y + 2)]^{y-1} dy + o(\tilde{c}^{2\nu-1}), \quad (\text{E.13})$$

where the integral is finite if  $\nu < \frac{1}{2}$ . Thus,  $N_1(1, \tilde{c}) + N_1(-1, \tilde{c}) = o(\tilde{c}^{2\nu-1})$  if  $\nu$  is strictly smaller than  $\frac{1}{2}$ .

When  $\nu = \frac{1}{2}$ , the integral in Eq. (E.13) is not defined, which requires different approach for this case. To extract the leading behavior, we performed integration by parts such that

$$\begin{aligned} \frac{N_1(\sigma, \tilde{c})}{2\sigma\nu^2\alpha(1-\alpha)} &= -2 \ln(\sqrt{2\tilde{c}}) \\ &+ 2 \int_0^{\infty} dy \ln(\sqrt{y} + \sqrt{y + 2\tilde{c}}) e^{-\zeta_{11}(y+2u(\sigma)\tilde{c}, \sigma, \tilde{c})} \frac{d\zeta_{11}}{dy} \\ &= -\ln(2\tilde{c}) + 2 \int_0^{\infty} \frac{e^{-\sqrt{x}}}{2\sqrt{x}} \ln(2\sqrt{x}) dx + o(1) \\ &= -\ln \tilde{c} + \ln 2 - 2\gamma + o(1), \end{aligned} \quad (\text{E.14})$$

where we have used ( $S > 0$ )

$$\frac{d}{dx} \ln(\sqrt{x} + \sqrt{x + S}) = \frac{1}{2\sqrt{x(x + S)}}, \quad (\text{E.15})$$

and  $\gamma \approx 0.5771$  is the Euler-Mascheroni constant. Still,  $N_1(1, \tilde{c}) + N_1(-1, \tilde{c}) = o(1)$  even for  $\nu = \frac{1}{2}$ . Hence,  $N_1$  does not contribute to the leading behavior of  $\psi(\tilde{c})$ .

Now we move on to the analysis of  $N_2$ . We first write  $N_2$  as

$$\begin{aligned} \frac{N_2(\sigma, \tilde{c})}{\sigma\omega_{\sigma}\nu^2} &= \int_{-\infty}^{-u(\sigma)\tilde{c}} [x(x + \sigma\tilde{c})]^{y-1} e^{-\zeta_{21}(-x, \sigma, \tilde{c})} dx \\ &= \int_0^{\infty} [y(y + \tilde{c})]^{y-1} e^{-\zeta_{21}(y+u(\sigma)\tilde{c}, \sigma, \tilde{c})} dy, \end{aligned} \quad (\text{E.16})$$

where we have made a change of variables  $y = -x - u(\sigma)\tilde{c}$  and

$$\begin{aligned} \zeta_{21}(z, \sigma, \tilde{c}) &= (1 - \omega_{\sigma}) [z^{\nu} + \ln K(-z + 2\sigma\tilde{c})] \\ &+ z^{\nu} + (z - \sigma\tilde{c})^{\nu} + e^{-z^{\nu}} - e^{-(z + \sigma\tilde{c})^{\nu}} \end{aligned} \quad (\text{E.17})$$

Note that  $\ln K(z) = 0$  if  $z > 0$ . As above, the change of variables to  $z = y/\tilde{c}$  gives

$$\frac{N_2(\sigma, \tilde{c})}{\sigma \omega_\sigma v^2} = \tilde{c}^{2\nu-1} \int_0^\infty dz [z(z+1)]^{\nu-1} e^{-\zeta_{22}(z, \sigma, \tilde{c})}, \quad (\text{E.18})$$

where  $\zeta_{22}(z, \sigma, \tilde{c}) = \zeta_{21}(\tilde{c}z + \tilde{c}u(\sigma), \sigma, \tilde{c})$  with the property  $\zeta_{22}(z, \sigma, 0) = 0$ . Thus the leading behavior of  $N_2$  is

$$\begin{aligned} N_2(\sigma, \tilde{c}) &\approx \sigma \omega_\sigma v^2 \tilde{c}^{2\nu-1} \int_0^\infty dz [z(z+1)]^{\nu-1} \\ &= \sigma \omega_\sigma v^2 \tilde{c}^{2\nu-1} \frac{\Gamma(1-2\nu)\Gamma(\nu)}{\Gamma(1-\nu)}, \end{aligned} \quad (\text{E.19})$$

which is valid for  $\nu < \frac{1}{2}$ . For  $\nu = \frac{1}{2}$ , integrating by parts gives

$$\begin{aligned} \frac{N_2(\sigma, \tilde{c})}{\sigma \omega_\sigma v^2} &= -\ln \tilde{c} \\ &+ 2 \int_0^\infty dy \ln \left[ \sqrt{y} + \sqrt{y + \tilde{c}} \right] e^{-\zeta_{21}(y+u(\sigma)\tilde{c}, \sigma, \tilde{c})} \frac{d\zeta_{21}}{dy} \\ &= -\ln \tilde{c} + 2 \int_0^\infty dy \frac{\ln(2\sqrt{y})}{\sqrt{y}} e^{-2\sqrt{y}} + o(1) \\ &= -\ln \tilde{c} - 2\gamma + o(1). \end{aligned} \quad (\text{E.20})$$

Since  $\omega_1 - \omega_{-1} = 2\alpha - 1$ , we finally get the leading behavior of  $\psi(\tilde{c})$  for  $\nu < \frac{1}{2}$  as

$$\begin{aligned} \psi(\tilde{c}) &\approx (2\alpha - 1)v^2 \tilde{c}^{2\nu-1} \frac{\Gamma(1-2\nu)\Gamma(\nu)}{\Gamma(1-\nu)}, \\ \int_0^{\tilde{c}} \psi(x) dx &= (2\alpha - 1)\tilde{c}^{2\nu} \frac{\Gamma(1-2\nu)\Gamma(\nu+1)}{2\Gamma(1-\nu)}, \end{aligned} \quad (\text{E.21})$$

and for  $\nu = \frac{1}{2}$  as

$$\begin{aligned} \psi(\tilde{c}) &\approx \frac{2\alpha - 1}{4} (-\ln \tilde{c} - 2\gamma), \\ \int_0^{\tilde{c}} \psi(x) dx &= -\frac{2\alpha - 1}{4} \tilde{c} \ln(e^{2\gamma-1} \tilde{c}). \end{aligned} \quad (\text{E.22})$$

## References

- Aita, T., Uchiyama, H., Inaoka, T., Nakajima, M., Kokubo, T., Husimi, Y., 2000. Analysis of a local fitness landscape with a model of the rough mt. fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopoly.* 54, 64–79.
- Bank, C., Hietpas, R. T., Wong, A., Bolon, D. N., Jensen, J D., 2014. A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. *Genetics* 196, 841–852.
- de Haan, L., Ferreira, A., 2006. *Extreme Value Theory*. Springer.
- de Visser, J. A. G. M., Krug, J., 2014. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* 15, 480–490.
- Desai, M. M., Fisher, D. S., 2007. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics* 176, 1759–1798.
- Flyvbjerg, H., Lautrup, B., 1992. Evolution in a rugged fitness landscape. *Phys. Rev. A* 46, 6714–6723.
- Franke, J., Klözer, A., de Visser, J. A. G. M., Krug, J., 2011. Evolutionary accessibility of mutational pathways. *PLoS Comp. Biol.* 7, e1002134.
- Franke, J., Wergen, G., Krug, J., 2010. Records and sequences of records from random variables with a linear trend. *J. Stat. Mech.:Theory Exp.*, P10013.
- Gerrish, P. J., Lenski, R. E., 1998. The fate of competing beneficial mutations in an asexual population. *Genetica* 102–103, 127–144.
- Gillespie, J. H., 1983. A simple stochastic gene substitution model. *Theor. Popul. Biol.* 23, 202–215.
- Gillespie, J. H., 1984. Molecular evolution over the mutational landscape. *Evolution* 38, 1116–1129.
- Hegarty, P., Martinsson, A., 2014. On the existence of accessible paths in various models of fitness landscapes. *Adv. Appl. Prob.* 24, 1375–1395.
- Hill, W. G., Robertson, A., 1966. The effect of linkage on the limits to artificial selection. *Genet. Res. Camb.* 8, 269–294.
- Jain, K., 2011. Number of adaptive steps to a local fitness peak. *Europhys. Lett.* 96, 58006.
- Joyce, P., Rokyta, D. R., Beisel, C. J., Orr, H. A. 2008. A General Extreme Value Theory Model for the Adaptation of DNA Sequences Under Strong Selection and Weak Mutation. *Genetics* 180: 1627–1643
- Kauffman, S., Levin, S., 1987. Towards a general-theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* 128, 11–45.
- Kingman, J. F. C., 1978. A simple model for the balance between selection and mutation. *J. Appl. Prob.* 15, 1–12.
- Koekoek, R., Lesky, P., Swarttouw, R., 2010. *Hypergeometric Orthogonal Polynomials and Their  $q$ -Analogues*. Springer.
- Macken, C. A., Perelson, A. S., 1989. Protein evolution on rugged landscapes. *Proc. Nat. Acad. Sci. USA* 86, 6191–6195.
- Miller, C. R., Joyce, P., Wichman, H. A., 2011. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* 187, 185–202.
- Neidhart, J., 2014. *Fitness Landscapes, Adaptation and Sex on the Hypercube*. PhD theiss, University of Cologne (Full text is available at <http://kups.ub.uni-koeln.de/5878/>).



- Neidhart, J., Krug, J., 2011. Adaptive walks and extreme value theory. *Phys. Rev. Lett.* 107, 178102.
- Neidhart, J., Szendro, I. G., Krug, J., 2013. Exact Results for Amplitude Spectra of Fitness Landscapes. *J. Theor. Biol.* 332, 2018–227.
- Neidhart, J., Szendro, I. G., Krug, J., 2014. Adaptation in tunably rugged fitness landscapes: The Rough Mount Fuji Model. *Genetics* 198, 699–721.
- Nowak, S., Krug, J., 2015. Analysis of adaptive walks on nk fitness landscapes with different interaction schemes. *J. Stat. Mech.:Theory Exp.*, P06014.
- Orr, H. A., 2002. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* 56, 1317–1330.
- Orr, H. A., 2003. A minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.* 220, 241–247.
- Orr, H. A., 2003. The genetic theory of adaptation: A brief history. *Nat. Rev. Genet.* 6, 119–127.
- Orr, H. A., 2006. The population genetics of adaptation on correlated fitness landscapes: the block model. *Evolution* 60, 1113–1124.
- Orr, H. A., 2010. The population genetics of beneficial mutations. *Phil. Trans. R. Soc. B* 365, 1195–1201.
- Park, S.-C., Krug, J., 2007. Clonal interference in large populations. *Proc. Nat. Acad. Sci. USA* 104, 18135–18140.
- Park, S.-C., Krug, J., 2008. Evolution in random fitness landscapes: the infinite sites model. *J. Stat. Mech.:Theory Exp.*, P04014.
- Park, S.-C., Simon, D., Krug, J., 2010. The speed of evolution in large asexual populations. *J. Stat. Phys.* 138, 381–410.
- Park, S.-C., Szendro, I. G., Neidhart, J., Krug, J., 2015. Phase transition in random adaptive walks on correlated fitness landscapes. *Phys. Rev. E* 91, 042707.
- Perelson, A. S., Macken, C. A., 1995. Protein evolution on partially correlated landscapes. *Proc. Nat. Acad. Sci. USA* 92, 9657–9661.
- Rokyta, D. R., Beisel, C. J., Joyce, P., Ferris, M. T., Burch, C. L., Wichman, H. A., 2008. Beneficial Fitness Effects Are Not Exponential for Two Viruses. *J. Mol. Evol.* 67, 368–376.
- Rosenberg, N. A., 2005. A sharp minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.* 237, 17–22.
- Schenk, M., Szendro, I. G., Krug, J., de Visser, J. A. G. M., 2012. Quantifying the Adaptive Potential of an Antibiotic Resistance Enzyme. *PLoS Genetics* 8, e1002783.
- Schoustra, S. E., Bataillon, T., Gifford, D. R., Kassen, R., 2009. The Properties of Adaptive Walks in Evolving Populations of Fungus. *PLoS Biology* 7, e1000250.
- Seetharaman, S., Jain, K., 2011. Multiple adaptive substitutions during evolution in novel environments. *Genetics* 189, 1029–1043.
- Seetharaman, S., Jain, K., 2014. Adaptive walks and distribution of beneficial fitness effects. *Evolution* 68, 965–975.
- Szendro, I. G., Franke, J., de Visser, J. A. G. M., Krug, J., 2013a. Predictability of evolution depends nonmonotonically on population size. *Proc. Nat. Acad. Sci. USA* 110, 571–576.
- Szendro, I. G., Schenk, M. F., Franke, J., Krug, J., de Visser, J. A. G., 2013b. Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech.:Theory Exp.*, P01005.
- Weinberger, E. D., 1991. Local properties of Kauffman’s N-k model: A tunably rugged energy landscape. *Phys. Rev. A* 44, 6399–6413.
- Weinreich, D. M., Watson, R. A., Chao, L., 2005. Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59, 1165–1174.
- Wilke, C. O., 2004. The speed of adaptation in large asexual populations. *Genetics* 167, 2045–2053.
- Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. 6th Int. Cong. Genet.* 1, 356–366.